USFD2: Annotating Temporal Expresions and TLINKs for TempEval-2

Leon Derczynski
Dept of Computer Science
University of Sheffield
Regent Court
211 Portobello
Sheffield S1 4DP, UK
leon@dcs.shef.ac.uk

Robert Gaizauskas
Dept of Computer Science
University of Sheffield
Regent Court
211 Portobello
Sheffield S1 4DP, UK
robertg@dcs.shef.ac.uk

Abstract

We describe the University of Sheffield system used in the TempEval-2 challenge, USFD2. The challenge requires the automatic identification of temporal entities and relations in text.

USFD2 identifies and anchors temporal expressions, and also attempts two of the four temporal relation assignment tasks. A rule-based system picks out and anchors temporal expressions, and a maximum entropy classifier assigns temporal link labels, based on features that including descriptions of associated temporal signal words. USFD2 identified temporal expressions successfully, and correctly classified their type in 90% of cases. Determining the relation between an event and time expression in the same sentence was performed at 63% accuracy, the second highest score in this part of the challenge.

1 Introduction

The TempEval-2 (Pustejovsky and Verhagen, 2009) challenge proposes six tasks. Our system tackles three of these: task A – identifying time expressions, assigning `TIMEX3` attribute values, and anchoring them; task C – determining the temporal relation between and event and time in the same sentence; and task E – determining the temporal relation between two main events in consecutive sentences. For our participation in the task, we decided to employ both rule- and ML-classifier-based approaches. Temporal expressions are dealt with by sets of rules and regular expressions, and relation labelling performed by NLTK's[1] maximum entropy classifier with rule-based processing applied during feature generation. The features (described in full in Section 2) included at-

---

[1]See http://www.nltk.org/ .

tributes from the TempEval-2 training data annotation, augmented by features that can be directly derived from the annotated texts. There are two main aims of this work; to create a rule-based temporal expression annotator that included knowledge from work published since GUTime (Mani and Wilson, 2000) and measure its performance, and to measure the performance of a classifier that includes features based on temporal signals.

Our entry to the challenge, USFD2, is a successor to USFD (Hepple et al., 2007). In the rest of this paper, we will describe how USFD2 is constructed (Section 2), and then go on to discuss its overall performance and the impact of some internal parameters on specific TempEval tasks. Regarding classifiers, we found that despite using identical feature sets across relation classification tasks, performance varied significantly. We also found that USFD2 performance trends with TempEval-2 did not match those seen when classifiers were trained on other data while performing similar tasks. The paper closes with comments about future work.

2 System Description

The TempEval-2 training and test sets are partitioned into data for entity recognition and description, and data for temporal relation classification. We will first discuss our approach for temporal expression recognition, description and anchoring, and then discuss our approach to two of the relation labelling tasks.

2.1 Identifying, describing and anchoring temporal expressions

Task A of TempEval-2 requires the identification of temporal expressions (or **timex**s) by defining a start and end boundary for each expression, and assigning an ID to it. After this, systems should attempt to describe the temporal expression, determining its type and value (described below).

Our timex recogniser works by building a set of n-grams from the data to be annotated ($1 \leq n \leq 5$), and comparing each n-gram against a hand-crafted set of regular expressions. This approach has been shown to achieve high precision, with recall increasing in proportion to ruleset complexity (Han et al., 2006; Mani and Wilson, 2000; Ahn et al., 2005). The recogniser chooses the largest possible sequence of words that could be a single temporal expression, discarding any sub-parts that independently match any of our set of regular expressions. The result is a set of boundary-pairs that describe temporal expression locations within documents. This part of the system achieved 0.84 precision and 0.79 recall, for a balanced f1-measure of 0.82.

The next part of the task is to assign a type to each temporal expression. These can be one of TIME, DATE, DURATION, or SET. USFD2 only distinguishes between DATE and DURATION timexs. If the words *for* or *during* occur in the three words before the timex, the timex ends with an *s* (such as in *seven years*), or the timex is a bigram whose first token is *a* (e.g. in *a month*), then the timex is deemed to be of type DURATION; otherwise it is a DATE. These three rules for determining type were created based on observation of output over the test data, and are correct 90% of the time with the evaluation data.

The final part of task A is to provide a value for the timex. As we only annotate DATEs and DURATIONs, these will be either a fixed calendrical reference in the format YYYY-MM-DD, or a duration in according to the TIMEX2 standard (Ferro et al., 2005). Timex strings of *today* or *now* were assigned the special value PRESENT_REF, which assumes that *today* is being used in a literal and not figurative manner, an assumption which holds around 90% of the time in newswire text (Ahn et al., 2005) such as that provided for TempEval-2. In an effort to calculate a temporal distance from the document creation time (DCT), USFD2 will then check to see if numeric words (e.g. *one*, *seven hundred*) are in the timex, as well as words like *last* or *next* which determine temporal offset direction. This distance figure supplies either the second parameter to a DURATION value, or helps calculate DCT offset. Strings that describe an imprecise amount, such as *few*, are represented in duration values with an X, as per the TIMEX2 standard. We next search the

Table 1: Features used by USFD2 to train a temporal relation classifier.

Feature	Type
For events	
Tense	String
Aspect	String
Polarity	pos or neg
Modality	String
For timexs	
Type	Timex type
Value	String
Describing signals	
Signal text	String
Signal hint	Relation type
Arg 1 before signal?	Boolean
Signal before Arg 2?	Boolean
For every relation	
Arguments are same tense	Boolean
Arguments are same aspect	Boolean
Arg 1 before Arg 2?	Boolean
For every interval	
Token number in sentence / 5	Integer
Text annotated	String
Interval type	event or timex

timex for temporal unit strings (e.g. *quarter*, *day*). This helps build either a duration length or an offset. If we are anchoring a date, the offset is applied to DCT, and date granularity adjusted according to the coarsest temporal primitive present – for example, if DCT is 1997-06-12 and our timex is *six months ago*, a value of 1997-01 is assigned, as it is unlikely that the temporal expression refers to the day precisely six months ago, unless followed by the word *today*.

Where weekday names are found, we used Baldwin's 7-day window (Baldwin, 2002) to anchor these to a calendrical timeline. This technique has been found to be accurate over 94% of the time with newswire text (Mazur and Dale, 2008). Where dates are found that do not specify a year or a clear temporal direction marker (e.g., *April 17* vs. *last July*), our algorithm counts the number of days between DCT and the next occurrence of that date. If this is over a limit f, then the date is assumed to be last year. This is a very general rule and does not take into account the tendency of very-precisely-described dates to be closer to DCT, and far off dates to be loosely specified. An f of 14 days gives the highest performance based on the TempEval-2 training data.

Anchoring dates / specifying duration lengths was the most complex part of task A and our naïve rule set was correct only 17% of the time.

Table 2: A sample of signals and the TempEval-2 temporal relation they suggest.

Signal phrase	Suggested relation
previous	AFTER
ahead of	BEFORE
so far	OVERLAP
thereafter	BEFORE
in anticipation of	BEFORE
follows	AFTER
since then	BEFORE
soon after	AFTER
as of	OVERLAP-OR-AFTER
throughout	OVERLAP

2.2 Labelling temporal relations

Our approach for labelling temporal relations (or **TLINK**s) is based on NLTK's maximum entropy classifier, using the feature sets initially proposed in Mani et al. (2006). Features that describe temporal signals have been shown to give a 30% performance boost in TLINKs that employ a signal (Derczynski and Gaizauskas, 2010). Thus, the features in Mani et al. (2006) are augmented with those used to describe signals detailed in Derczynski and Gaizauskas (2010), with some slight changes. Firstly, as there are no specific TLINK/signal associations in the TempEval-2 data, USFD2 needs to perform signal identification and then associate signals with a temporal relation or entity. Secondly, a look-up list is used to provide TLINK label hints based on a signal word. A list of features employed by USFD2 is in Table 1.

We used a simplified version of the approach in Cheng et al. (2007) to pick signal words. This involved hand-picking a list of signal phrases that occur in TimeBank (Pustejovsky et al., 2003) with a frequency of 2 or more, and associating a signal from this list with a temporal entity if it is in the same sentence and clause. The nearest signal is chosen in the case of conflict.

As this list of signal phrases only contained 42 entries, we also decided to define a "most-likely" temporal relation for each signal. This was done by imagining a short sentence of the form *event1 – signal – event2*, and describing the type of relation between event 1 and event 2. An excerpt from these entries is shown in Table 2. The hint from this table was included as a feature. Determining whether or not to invert the suggested rela-

tion type based on word order was left to the classifier, which is already provided with word order features. It would be possible to build these suggestions from data such as TimeBank, but a number of problems stand in the way; the TimeML and TempEval-2 relation types are not identical, word order often affects the actual relationship type suggested by a signal (e.g. compare *He ran home before he showered* and *Before he ran home, he showered*), and noise in mined data is a problem with the low corpus occurrence frequency of most signals.

This approach was used for both the intra-sentence timex/event TLINK labelling task and also the task of labelling relations between main events in adjacent sentences.

3 Discussion

USFD2's rule-based element for timex identification and description performs well, even achieving above-average recall despite a much smaller rule set than comparable and more complex systems. However, the temporal anchoring component performs less strongly. The "all-or-nothing" metric employed for evaluating the annotation of timex values gives non-strict matches a zero score (e.g. if the expected answer is 1990-05-14, no reward is given for 1990-05) even if values are close, which many were.

Unlike previous approaches using a maximum entropy classifier and comparable feature set (Mani et al., 2006; Derczynski and Gaizauskas, 2010), the classification of event-event relations was less accurate than the classification of event-timex relations. Event-event relation classification accuracy was good (63%), despite the lack of explicit signal/TLINK associations and a sophisticated signal recognition and association mechanism. This is signficantly higher than USFD2's accuracy in task E – 45% – though the latter is a harder task, as most TempEval-2 systems performed significantly worse at this task than event/timex relation classification.

Signal information was not relied on by many TempEval 2007 systems (Min et al. (2007) discusses signals to some extent but the system described there only includes a single feature – the signal text), and certainly no processing of this data was performed for that challenge. USFD2 begins to leverage this information, and gives very competitive performance at event/timex classifi-

cation. Figures on the impact that signal-based features had on our system's performance will be available shortly.

The performance of classifier based approaches to temporal link labelling seems to be levelling off – the 60%‑70% relation labelling accuracy of work such as Mani et al. (2006) has not been greatly exceeded. This performance level is still the peak of the current generation of systems. Recent improvements, while employing novel approaches to the task that rely on interactions between individual links or non-trivial linguistic information, still yield marginal improvements (e.g. Yoshikawa et al. (2009)). It seems that to break through this performance "wall", we need to continue to innovate with and discuss temporal relation labelling, drawing on information and knowledge from many sources to build practical high-performance systems.

4   Conclusion

In this paper, we have presented USFD2, a novel system that annotates temporal expressions and temporal links in text. The relies on new handcrafted rules, existing rule sets, machine learning and temporal signal information to make its decisions. Although some of the TempEval-2 tasks are difficult, USFD2 manages to create good and useful annotations of temporal information. USFD2 is available via Google Code[2].

Acknowledgments

References

D. Ahn, S.F. Adafre, and MD Rijke. 2005. Towards task-based temporal extraction and recognition. In *Dagstuhl Seminar Proceedings*, volume 5151.

J.A. Baldwin. 2002. *Learning temporal annotation of French news*. Ph.D. thesis, Georgetown University.

Y. Cheng, M. Asahara, and Y. Matsumoto. 2007. Naist. japan: Temporal relation identification using dependency parsed tree. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 245–248. ACL.

L. Derczynski and R. Gaizauskas. 2010. Using signals to improve automatic classification of temporal relations. University of Sheffield. Submitted to ESSLLI 2010.

L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. Tides 2005 standard for the annotation of temporal expressions.

B. Han, D. Gates, and L. Levin. 2006. From language to time: A temporal expression anchorer. In *Temporal Representation and Reasoning, 2006. TIME 2006. Thirteenth International Symposium on*, pages 196–203.

M. Hepple, A. Setzer, and R. Gaizauskas. 2007. USFD: preliminary exploration of features and classifiers for the TempEval-2007 tasks. In *Proceedings of SemEval-2007*, pages 438–441.

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on ACL*, pages 69–76. ACL.

I. Mani, M. Verhagen, B. Wellner, C.M. Lee, and J. Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, page 760. ACL.

P. Mazur and R. Dale. 2008. Whats the date? High accuracy interpretation of weekday. In *22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK*, pages 553–560.

C. Min, M. Srikanth, and A. Fowler. 2007. LCC-TE: a hybrid approach to temporal relation identification in news text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 219–222. ACL.

J. Pustejovsky and M. Verhagen. 2009. SemEval-2010 task 13: evaluating events, time expressions, and temporal relations (TempEval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116. ACL.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003. The timebank corpus. In *Corpus Linguistics*, volume 2003, page 40.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *ACL-IJCNLP '09: Proceedings of 47th Annual Meeting of the ACL*, pages 405–413, Morristown, NJ, USA. ACL.

---

[2]See http://code.google.com/p/usfd2/ .