À la Recherche du Temps Perdu: extracting temporal relations from medical text in the 2012 i2b2 NLP challenge

Colin Cherry, Xiaodan Zhu, Joel Martin, Berry de Bruijn

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001624).

Information and Communication Technologies, National Research Council Canada, Ottawa, Ontario, Canada

Correspondence to
Dr Berry de Bruijn, Information and Communication Technologies, National Research Council Canada, 1200 Montreal Road, Building M50, Ottawa, ON, Canada K1A 0R6; berry.debruijn@nrc-cnrc.gc.ca

Received 3 January 2013
Revised 12 February 2013
Accepted 17 February 2013
Published Online First 23 March 2013

ABSTRACT

Objective An analysis of the timing of events is critical for a deeper understanding of the course of events within a patient record. The 2012 i2b2 NLP challenge focused on the extraction of temporal relationships between concepts within textual hospital discharge summaries.

Materials and methods The team from the National Research Council Canada (NRC) submitted three system runs to the second track of the challenge: typifying the time-relationship between pre-annotated entities. The NRC system was designed around four specialist modules containing statistical machine learning classifiers. Each specialist targeted distinct sets of relationships: local relationships, 'sectime'-type relationships, non-local overlap-type relationships, and non-local causal relationships.

Results The best NRC submission achieved a precision of 0.7499, a recall of 0.6431, and an F1 score of 0.6924, resulting in a statistical tie for first place. Post hoc improvements led to a precision of 0.7537, a recall of 0.6455, and an F1 score of 0.6954, giving the highest scores reported on this task to date.

Discussion and conclusions Methods for general relation extraction extended well to temporal relations, and gave top-ranked state-of-the-art results. Careful ordering of predictions within result sets proved critical to this success.

INTRODUCTION

The 2012 i2b2 NLP challenge, the sixth in its series, revolved around extraction of temporal relations from clinical text. It was designed as a two-stage process: (1) extraction and formalization of clinical and time concepts from the text, and (2) pairwise linking of concepts using BEFORE/OVERLAP/AFTER characterizations. The i2b2 challenge design mimicked that of the TempEval challenge,[1] part of SemEval-2007.

The current article presents the contribution of the National Research Council Canada (NRC) to the 2012 challenge. It briefly outlines the challenge's evaluation methodology, and then discusses in turn the components of the NRC system, system performance, limitations of the system, limitations of evaluation, and possible future directions for this type of research.

BACKGROUND AND SIGNIFICANCE

Temporal reasoning is essential for gaining a deeper understanding of the course of events described in a patient record, notably as a way to help establish causal relationships between events. Zhou and Hripcsak[2] provide an excellent overview of temporal reasoning with medical data, highlighting the complexities and rationales, and listing related research. Sun *et al*[3] further discuss the motivation for this research in their overview paper in this issue of the journal.

Within the i2b2 challenge, we decided to concentrate on subtask (2)—pairwise linking of concepts. We follow the approach of Mani *et al*,[4] where statistical classifiers are used to determine the temporal relation (if any) that links each pair of entities. As such, we draw heavily from the literature on non-temporal relation extraction for clinical documents.[5 6] Our work results in two primary contributions. First, we demonstrate that by dividing relationships into groups to be handled by specialist modules, techniques for non-temporal relation extraction can produce state-of-the-art results on the temporal relation extraction task. Second, our results highlight the importance of careful ordering of the resulting relationship list, and we call into question whether having an order-dependent evaluation is desirable.

METHOD

Data

A total of 190 training documents were made available. Each document was a free-text discharge summary for one patient's hospital stay. Text had been pre-split into sentences and pre-tokenized into words. Section headings were not explicitly marked, but could be inferred. On average, a document contained about 28 sentences (median; mean: 33, range: 3–162) and 418 words (median; mean: 502, range: 50–1752). The test set document collection (n=120) was designed to be entirely comparable, but appeared to contain slightly longer documents on average, with 38 lines per document (median; mean: 41, range: 4–128) and 568 words per document (median; mean: 657, range: 46–2248).

In addition to the textual material, files with concept annotations were made available to participants in subtask (2). There were two main types of concepts or 'extents': Events and Times (see figure 1). An Event can be an occurrence, problem, treatment, test, evidence, or clinical department. This typification is embedded in the annotation. Likewise, Times can be of type frequency, duration, date, or time. If applicable, annotations for Times and Events contained additional attributes for modality, polarity, and value. Per text, an average of 12 Time concepts and 87 Event-type extents were seen. For further details on the data and annotations, see Sun *et al*.[3]

To cite: Cherry C, Zhu X, Martin J, *et al*. *J Am Med Inform Assoc* 2013;**20**: 843–848.

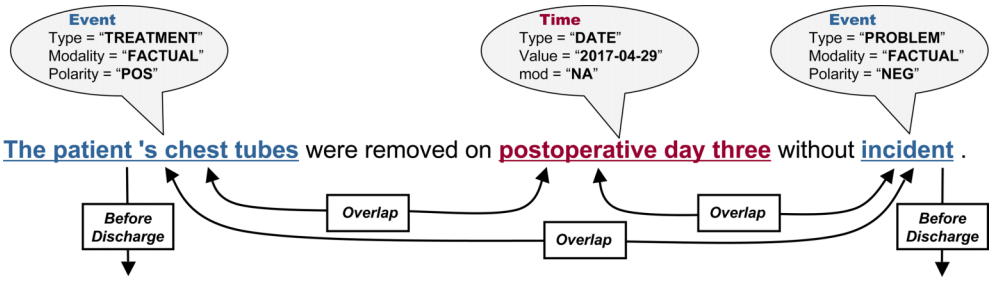

Figure 1 Example of a sentence from a discharge text, the provided annotations (balloons), and desired temporal relations (arrows).

Task

The task is to annotate pairs of extents as being in a time-relationship of type BEFORE, OVERLAP, or AFTER. These relationships are referred to as TLINKs. While temporal relationships could be assigned to each pair of extents, that would result in great redundancy. If it is known that A happened before B and B before C, then A-before-C is redundant. Stating it explicitly is unnecessary, and doing so is not rewarded in evaluation. To guide the development of systems, ground truth linkage data were provided for the training documents.

Evaluation

Sets of extracted relations are minimized (per document) to eliminate redundancy, and then compared to the minimal non-redundant ground truth set of relations. Matches between predicted relations and the ground truths are then summarized in precision, recall, and F1 score. UzZaman and Allen[7] describe this evaluation method in more detail.

SYSTEM

Our system is built using machine-learning components, which learn from patterns in the training material and then apply these learned patterns to the test material. We employed a modular system where four separate modules each predict links and thus contribute to a final set of links for the predicted time graph: (1) a local-link specialist, for links between extents within the same sentence; (2) a 'sectime' specialist, for links between extents in the text and the strictly defined time points for Admission and Discharge; (3) a non-local specialist to predict OVERLAP-type links between extents that occur further apart in the text, especially those that co-refer to the same concept; and (4) a high-precision rule-based specialist that finds links known (or discovered) to be causal in nature.

Additionally, two central modules provide essential support to these link-predicting modules: (1) a sorter module that receives the predictions from the individual specialists and merges them into an optimally tuned final list of predictions, and (2) a temporal reasoning component. The temporal reasoning component can both expand and contract a set of time-relation links. Expansion allows for a larger coverage (or closure) of links between extents and therefore a growth in meaningful training material. Contraction allows for properly measuring the non-redundant contribution of (sets of) temporal links. Also, the temporal reasoning component serves to identify contradictions in sets of relationships.

Throughout, development was guided by 10-fold cross-validation; randomized assignment of documents to folds was done once and remained unchanged for the various components. In the paragraphs below, we describe the temporal reasoning module, the four specialized prediction modules, and the prediction sorter module.

Temporal reasoning module

The space of possible TLINKs is very large, and the number of gold-standard TLINKs is relatively small. We address this imbalance by training all our machine-learned TLINK detectors on the transitive closure of the gold-standard graph.[4] This has the benefit of both reducing the prevalence of the majority class ('no link'), as well as reducing the noise caused by true relations that were missing from the ground truth.

Creating the transitive closure is a form of temporal inference. With only BEFORE, AFTER, and OVERLAP type relations, this potentially complex task[8] has been highly simplified. We represent a collection of TLINKs as a simple directed graph with labeled arcs. BEFORE relations are inverted into AFTER relations, giving us only two relation types in the graph: AFTER and OVERLAP. Our graph has one node per extent, with each AFTER TLINK represented as a directed From→To arc labeled AFTER. Since the OVERLAP relation is reflexive, each OVERLAP is represented as two directed arcs, From→To and To→From, both labeled OVERLAP. With this representation in place, we can carry out temporal inference through the transitive closure of the directed graph, using table 1 to determine the relation connecting A and C from the relations connecting A, B and B, C.

We use a simple arc-incremental variant of the Floyd–Warshall algorithm[9] to carry out transitive closure. Mimicking the organizer's evaluation tool, we resolve contradictions during incremental closure. TLINKs are added one by one in the order they appear in a file. If the addition of any arc would result in the inference of a TLINK 'A AFTER A,' we declare the arc as contradictory, and omit it from the closure. Thus, closure resolves contradictions as it infers new relations. We may also be interested in finding the minimal set of links that can infer a given input set. We can carry out greedy graph minimization during incremental closure (as is done by the evaluation tool), or we can conduct a full minimization through transitive reduction using the algorithm described by Aho *et al*.[10] We use our minimizations (both greedy and full) only to analyze our system's output and to critique the official evaluation tool. Our predicted relations are actually presented to the evaluation tool as is, without minimization.

The local-link specialist

The local-link specialist is a maximum entropy (ME) classifier that typifies links between extents within the same sentence.

Table 1 Extending links through inference

	B AFTER C	B OVERLAP C
A AFTER B	A AFTER C	A AFTER C
A OVERLAP B	A AFTER C	A OVERLAP C

For such links, we can employ very rich syntactic and semantic features. In addition, we use surface and structure features. Extents are processed in the order in which they are encountered in the text, and if necessary, gold-standard link types are reversed to reflect this (eg, 'B AFTER A' becomes 'A BEFORE B' if A (…) B was the order in the text). To balance the class distribution after temporal closure, we added two more copies of the negative ('no link') relations to the training data. This added 0.03 to our F1 score in cross-validation, using the provided evaluation tool.

We chose ME[11] because it showed as good performance as Support Vector Machine (SVM) and other classifiers in our prior research[5] and operated fast enough to cross-validate a wide variety of features in a limited development period. We used the ME package of OpenNLP.[12] We trained two different ME classifiers, one for event–event relations and one for event–time relations. We ignored all time–time relations since they are rare—201 out of 17 821 non-sectime relations in the training data, before closure expansion.

The feature engineering efforts in this task built upon our previous work in non-temporal relation extraction,[5] which achieved top performance in the 2010 i2b2 NLP challenge. The final set of features can be grouped into surface features, syntactic features, semantic features, and structural features. Supplementary online appendix A describes these in greater detail.

Surface features

Surface features describe how two extents are superficially linked within the text. These include word/extent n-grams, word/extent sequences, discretized counts of words, punctuation marks, and extents, as well as features that capture the order of extents. In general, we replaced words with extents where possible. For instance: given that 'multiple angioplasty stents' was known to be an event extent with the type 'treatment,' the phrase 'post multiple angioplasty stents' would yield a hybrid 2-gram ('post <event::treatment>') rather than a word 4-gram. Clinical narrative often lists events in conjunction, for example, 'the patient experienced episode of tongue biting, unresponsiveness, disorientation.' We introduced special handling for these situations, allowing our feature templates to skip over some extents when a lengthy conjunction separates a candidate extent pair. Specifically, we used rules to recognize and skip pieces of text that contain only commas, conjunctions, and extents.

Syntax features

Syntax features are derived from part-of-speech (POS) tags, as well as dependency trees. We first parsed the input text using Charniak's ME reranking parser[13] with its improved, self-trained biomedical parsing model.[14] The POS tags from these trees were used to create n-gram POS features analogous to the word n-gram features discussed above. Then, these constituent trees were converted into Stanford dependencies[15] and features were extracted to describe the structure of these trees. Specifically, we find the minimum-covering trees for each extent in a candidate relation, as well as the tree connecting the two extents. In cross-validation, syntax features added 0.017 to our F1 score.

Semantic features

The limited quantity of labeled data leads to a sparseness of features, which we attempt to reduce through smoothing by using both a manually created lexicon and a knowledge base. The lexicon consists of approximately 600 words and phrases that express the concepts of *before*, *after*, *causing*, *caused by*, *during*, *starting*, *continuing*, *ending*, *suddenly*, and *currently*. Supplementary online appendix B contains the entire lexicon. If a word or a phrase is in the lexicon, it is substituted with its subcategory label (eg, *causing*). We also used MetaMap[6 16] to map extents to the semantic categories in the Unified Medical Language System (UMLS), generating UMLS label pairs as features for candidate relations.

Structural features

Up to now, relations are classified independently. We also adapted the method proposed by Roberts *et al*[17] to consider prior decisions during classification, and it led to consistent gains during development. The temporal relations of closer extent pairs are predicted first and are then reused as features for predictions on the pairs that are farther apart.

Sectime specialist

The annotation standard for the i2b2 task specifies that each event in a document must be connected to either the admission or discharge date, via a 'sectime' link. An event in the patient history section links to the admission date, while an event in the hospital course section links to the discharge date. Because these sectime TLINKs are constrained by very specific rules, we target them using a special-purpose classifier.

We trained a multi-class SVM to categorize each event into one of seven classes: BeforeAdmission, OverlapsAdmission, AfterAdmission, BeforeDischarge, OverlapsDischarge, AfterDischarge, Empty. We selected an SVM classifier over ME, because it achieved higher accuracies in cross-validation, and because precision is high enough that we did not feel we would benefit from inducing probabilities over classes. Our SVM is an in-house implementation, similar to LIBLINEAR.[18] Cross-validated in isolation, this specialist achieved 89.6% accuracy by constructing six binary feature templates from four binary template atoms:

- Bias: A feature that is always on
- Loc: Indicates whether the event appears in the patient history section or the hospital course section. These two sections are separated by a 'hospital course' line, defined heuristically as the shortest line that matches the case-insensitive regular expression 'hospital( )+course.'
- LineNo: Indicates the line number of the event, divided into the following bins: 1, 2, 3, 4, 5, 6+.
- Str: Indicates the complete lowercased string corresponding to this event as it appears in the text.

The atoms were composed into the following six templates, where • indicates composition: Bias • Y, Loc • Y, LineNo • Y, Str • Y, Loc • LineNo • Y, Loc • Str • Y. Note that every feature is concatenated with the class Y to enable multi-class classification.

Non-local event-overlap specialist

Our only machine-learned classifier for non-local TLINKs is an event-to-event overlap specialist; it never assigns BEFORE or AFTER relations. We were motivated to build this classifier by a number of observations. Non-local TLINK predictions are extremely risky, as there are far more extent–extent pairs that could be linked than there are true links, even when considering the closure of the gold-standard links. Therefore, every attempted non-local link is likely to reduce precision. Also, due to the nature of the default evaluation, BEFORE and AFTER links only result in a recall gain if the proposed links are actually minimal.

Because of their non-reflexive nature, non-minimal BEFORE and AFTER links cannot be used to infer minimal ones. These relations are then a precision risk with little expected recall gain. On the other hand, OVERLAP is reflexive, making reverse inference possible. That is, several non-minimal OVERLAPS can be used to infer minimal ones. For example, take the minimal chain A=B=C=D; given the alternative links A=D, A=C, B=D, we can infer the entirety of the original chain. In contrast, none of A<B<C<D can be recovered from A<D, A<C, B<D. Therefore, OVERLAP bets are more likely to pay off in terms of recall.

We observed that many non-local OVERLAP relations are in fact well-understood co-reference relations between events. With this in mind, we constrained our search space to event pairs within five sentences and perfectly matching event attributes (type, modality, and polarity). This gave roughly 34 k training pairs, down from 999 k without this constraint, with a 17% a priori label probability for OVERLAP. We trained a binary ME classifier to assign each event pair to either the OVERLAP or Other class. Our features (see supplementary online appendix A) summarized the two events involved in terms of their local contexts and the words and extents that come between them. The resulting classifier built 2.5 M features from these templates. Unlike all of our attempts at training a BEFORE—OVERLAP—AFTER non-local TLINK classifier, the overlap specialist consistently improved the complete system.

Non-local rule-based specialist

Despite the risk of non-local BEFORE or AFTER links, we included a rule-based specialist for hypothesizing such links. This specialist applied rules to link substrings that appeared in two extents. Specifically, the rules were of the form: Extent-A is BEFORE/AFTER Extent-B if they both match or contain a specific trigger-(sub) string. For example, 'anesthetization' happens before 'resuscitation,' and 'shortness of breath' happens before 'nebulizers.' We only considered rules that were accurate in the training data. In particular, a rule was considered viable in the training data if it appeared with a frequency of 10 or higher and with a precision of 85% or higher. We filtered out accidental rules by requiring plausibility to a human annotator. For example, there is a simple explanation of why anaesthetization should occur before resuscitation in a hospital visit, and why injury happens before rehabilitation.

We wrapped these rules in a ME classifier by defining features for each rule, so the resulting links could be assigned probabilities. Unfortunately, as training-data precision was a major component in rule selection, the classifier's probabilities tended to be overly confident. For similar reasons, we were uncertain that cross-validation on the training data would give realistic estimates of rule performance. Because of this, we only included these rules in two of our three system submissions.

Prediction sorter module

Our predictions, as generated by the sectime specialist, the local specialist, the non-local overlap specialist, and the non-local rule-based specialist, are collated by a prediction sorter module. It optimizes the order in which TLINKs are presented to the official evaluation script, as we discovered this to have a significant impact on the final score. There are two reasons for this order-dependence.

First, the evaluation script carries out temporal reasoning through its own incremental transitive closure, as the closure of the set of system TLINKs is required to calculate recall. During closure, links are added by their order in the file, and if any link results in a contradiction, it is dropped from the closure. By listing high-confidence links first, they are more likely to contribute to the closure and improve recall.

Second, and less obviously, the order of TLINKs can affect the system's precision. The evaluation script carries out temporal minimization of the system TLINKs before calculating precision. Each link in the minimal system graph is checked against the closure of the gold standard, and precision is calculated as:

$$\frac{\#\text{ links in system min graph that are also in gold closure}}{\#\text{ links in system min graph}}$$

The minimization carried out by the evaluation is actually order-dependent. Links are added one by one, constructing the transitive closure incrementally. If a link to be added is already in the closure, it is left out of the minimized graph. The resulting graphs are not truly minimal. Imagine a minimal graph A<B<C<D, and an adversary that presents the ordered list of links A<D, A<C, B<D, C<D, B<C, A<B; the greedy minimization would contain all six links, as the three redundant links A<D, A<C, and B<D are placed before the links that make them redundant. Thus, placing high-precision links near the top of the file can improve precision as well as recall, as these links are less likely to be minimized away, allowing them to increment both the numerator and denominator when calculating precision. The larger denominator limits the impact of later precision errors. This provides further motivation to sort our links according to their confidence.

Our final submission sorted the system output as follows. Lacking probabilities, the SVM-driven sectimes are simply placed first, as they are very accurate. Following these are links from the remaining three components, sorted together according to their ME posterior probabilities. In general, links produced by rules tended to be near the top of the lists, while the non-local overlap specialist and the local specialist mingled more evenly. No correction was carried out for redundancy or contradiction, as either reduced our F1 score using the official evaluation tool.

RESULTS AND DISCUSSION

The challenge allowed submission of result sets from three systems. All three of our systems employed the same local specialist, sectime specialist, and non-local event-overlap specialist. The non-local rule-based specialist was not used in System 1, was used with a limited set of rules in System 2, and with an expanded set of rules in System 3. Table 2 shows that performance gains of rules during development did not carry over into the test phase. Our best test scores were achieved without any rules (System 1). Among all challenge submissions, it statistically tied for first place with Vanderbilt (F1 score of 0.6932). An alternative System 1, modified during post hoc analyses and

Table 2 Training set and test set performance of system submissions

	Training set (n=190)			Test set (n=120)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
System 1	0.7573	0.6499	0.6995	0.7499	0.6431	0.6924
System 2	0.7727	0.6543	0.7086	0.7474	0.6433	0.6914
System 3	0.7825	0.6567	0.7141	0.7273	0.6449	0.6837
Post-hoc	0.7686	0.6570	0.7085	0.7537	0.6455	0.6954

Table 3 Ablation test results

	Specialist tested alone			System without specialist			
Specialist	Precision	Recall	F1 score	Precision	Recall	F1 score	Delta F1
All	0.7273	0.6449	0.6837	0.7273	0.6449	0.6837	–
Sectime	0.9274	0.2661	0.4136	0.6541	0.3997	0.4962	−0.1875
Local	0.5960	0.3573	0.4468	0.8333	0.3005	0.4417	−0.2420
Non-local overlap	0.6166	0.0358	0.0676	0.7316	0.6142	0.6678	−0.0159
Non-local rules	0.7623	0.0030	0.0060	0.7499	0.6431	0.6924	0.0083

discussed later in this section, scored higher than both with an F1 score of 0.6954, and stands at this time as the highest reported score on this task. In document-by-document comparison, F1 scores are significantly higher for the post hoc system than for System 1 (two-tailed paired t test, $p<0.001$).

In table 3, we compare the four specialists according to their performance on the test set. As a basis, we take System 3, which contains all components. First, we evaluated each of our four components in isolation, then a system with that component taken out.

As these results show that our strongest and most important component is the local specialist, which impacts F1 by roughly 0.24, with the sectime specialist following close behind with an F1 impact of 0.19. The two non-local components play a minor role. The overlap specialist buys some recall at the cost of precision, with a small but significant benefit to the F1 score (two-tailed paired t test, $p<0.001$). Non-local rules brought improvements to precision on the training data, but this effect failed to materialize on the test data.

For each specialist, we also calculated the percentage of TLINKs that proved redundant after minimization in our cross-validation data: 64.4% of sectime TLINKs were redundant with links from the other components, as were 43.7% of local TLINKs, 39.4% of the overlap specialist TLINKs, and 59.6% of the rule TLINKs. It is interesting to note that the evaluation provides no incentive to add non-redundant, non-minimal links to the temporal graph; for example, if the gold standard contains A<B and B<C, no credit or penalty is given to a system that returns A<C. Such links add information, but because it is at the wrong scale, it is ignored in recall calculations.

As noted, the evaluation uses a greedy minimization; therefore, the order of TLINKs can affect the system's reported precision. To illustrate this effect, we reverse the order of TLINKS in the output of our System 1 on the test set (which had a precision of 0.7499, recall of 0.6431, and F1 score of 0.6924). The same set of predicted links—but now sorted in reverse order—causes the evaluation script to report a precision of 0.6443, recall of 0.6320, and F1 score of 0.6381, for a loss of 0.054 F1, due almost entirely to a 0.1 drop in precision.

Over the course of our post hoc analyses, we realized more fully how OVERLAP-type links carried lower precision/recall risk than BEFORE/AFTER links, so the prediction sorter module was altered to sort OVERLAP links ahead of all other non-sectime links. By changing the sort order alone, both precision and recall improved, and the F1 score increased for both the training set, from 0.6995 to 0.7085, and the test set, from 0.6924 to 0.6954 (table 2, bottom row).

Our analyzes suggest that in the future, it may be worthwhile to evaluate temporal relations in a manner that does not depend on the order of the system's link list. This would require the use of a complete minimization, as well as a method to handle contradictions in an order-independent manner.

CONCLUSION

For the 2012 i2b2 NLP challenge on temporal relation extraction, the team from National Research Council Canada developed an ensemble system consisting of a core of four prediction specialists. Our highest scoring system configuration, developed after the competition closed, gave a precision of 0.7537, recall of 0.6455, and F1 score of 0.6954, resulting in the highest reported score on this task to date. Our best submitted system resulted in an F1 score of 0.6924, a statistical tie for first place.

Each prediction specialist targeted a distinct subgroup of possible links, using machine learning classifiers trained on the expanded transitive closure of the gold-standard data. The sectime specialist concentrated solely on links between any extent in the text and the admission or discharge date of the report. The local specialist was motivated by the observed high link density between extents within the same sentence, and proved able to leverage syntactic relationships between extents to infer the temporal ones. By targeting co-reference-type links, the non-local overlap specialist only searched for OVERLAP-type links, which are more likely to pay off in terms of recall due to their reflexive nature. The non-local rule-based specialist would try to capture the BEFORE and AFTER links and tried to mitigate the inherent risk of including such links by forcing the rules to be of high precision.

An important finding was the effect that ordering had on the final scores, due to design decisions in the evaluation methodology. The very same set of results but sorted in reverse order, would cause a drop in the F1 score of no less than 0.054. It underlines that evaluation with complete minimization, combined with an alternative way to handle contradictions, would be invaluable to the field.

Contributors CC designed and implemented the temporal reasoning module and the sectime and non-local overlap specialist. XZ designed and implemented the local-link specialist. JM designed and implemented the non-local rule-based specialist. BdeB coordinated participation in the challenge and preparation of the manuscript. All authors contributed to the overall system design, creative and methodological decisions, and analysis of results. All authors contributed to and approved the paper.

Funding The i2b2 NLP challenge was supported by Informatics for Integrating Biology and the Bedside (i2b2) award number 2U54LM008748 from the NIH/National Library of Medicine (NLM), by the National Heart, Lung, and Blood Institute (NHLBI), and by award number 1R13LM01141101 from the NIH/NLM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, NHLBI, or the National Institutes of Health.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Verhagen M, Gaizauskas R, Schilder F, *et al*. SemEval-2007 task 15: Tempeval temporal relation identification. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*; June 2007, Prague, Czech Republic: Stroudsburg, PA: Assoc for Computational Linguistics, 2007:75–80. doi:10.3115/1621474.1621488

2 Zhou L, Hripcsak G. Temporal reasoning with medical data–a review with emphasis on medical natural language processing. *J Biomed Inform* 2007;40:183–202.

3 Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge overview. *J Am Med Inform Assoc* 2013;20:806–13.

4 Mani I, Verhagen M, Wellner B, *et al*. Machine learning of temporal relations. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia. Stroudsburg, PA: Association for Computational Linguistics, 2006:753–60.

5 de Bruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;18:557–62.

6 Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17:229–36.

7 UzZaman N, Allen J. Temporal evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR. Stroudsburg, PA: Association for Computational Linguistics, 2011:351–6.

8 Vilain M, Kautz H. Constraint propagation algorithms for temporal reasoning. The Fifth National Conference on Artificial Intelligence (AAAI-86); Los Altos, CA: Philadelphia, PA: Morgan Kaufmann, 1986:377–82.

9 Floyd RW. Algorithm 97: shortest path. *Commun ACM* 1962;5:345.

10 Aho A, Garey M, Ullman J. The transitive reduction of a directed graph. *SIAM J Comput* 1972;1:131–7.

11 Berger A, Pietra V, Pietra S. A maximum entropy approach to natural language processing. *Computational Linguistics* 1996;22:39–71.

12 http://opennlp.sourceforge.net/projects.html (updated Sep 23, 2010; accessed 12 July 2012).

13 Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2005:173–80. doi:10.3115/1219840.1219862

14 McClosky D, Charniak E, Johnson M. Effective self-training for parsing. In: Proceedings of Human Language Technology conference / North American Chapter of the Association for Computational Linguistics Annual Meeting. New York, NY, USA. Stroudsburg, PA: Association for Computational Linguistics, 2006:152–9. doi:10.3115/1220835.1220855

15 de Marneffe MC, MacCartney B, Manning CD. Generating Typed Dependency Parses from Phrase Structure Parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. Genoa, Italy. Paris, France: European Language Resources Association, 2006;449–54.

16 National Library of Medicine. UMLS. Bethesda, MD: U.S. National Library of Medicine http://www.nlm.nih.gov/research/umls/about_umls.html (updated Aug 29, 2011; accessed 27 Jun 2012).

17 Roberts K, Rink B, Harabagiu S. Extraction of Medical Concepts, Assertions, and Relations from Discharge Summaries for the Fourth i2b2/VA Shared Task Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data, Washington, DC, 2010. Boston, MA: Informatics for Integrating Biology and the Bedside (i2b2).

18 Fan R-E, Chang K-W, Hsieh C-J, *et al*. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008;9:1871–74.