Extraction des relations temporelles entre événements médicaux dans des comptes rendus hospitaliers

Pierre Zweigenbaum[1] Xavier TANNIER[1,2]
[1] LIMSI–CNRS, Orsay [2]Univ. Paris-Sud, Orsay

RÉSUMÉ
Le défi i2b2/VA 2012 était dédié à la détection de relations temporelles entre événements et expressions temporelles dans des comptes rendus hospitaliers en anglais. Les situations considérées étaient beaucoup plus variées que dans les défis TempEval. Nous avons donc axé notre travail sur un examen systématique de 57 situations différentes et de leur importance dans le corpus d'apprentissage en utilisant un oracle, et avons déterminé empiriquement le classifieur qui se comportait le mieux dans chaque situation, atteignant ainsi une F-mesure globale de 0,623.

ABSTRACT
Extraction of temporal relations between clinical events in clinical documents

The 2012 i2b2/VA challenge focused on the detection of temporal relations between events and temporal expressions in English clinical texts. The addressed situations were much more diverse than in the TempEval challenges. We thus focused on the systematic study of 57 distinct situations and their importance in the training corpus by using an oracle, and empirically determined the best performing classifier for each situation, thereby achieving a 0.623 F-measure.

MOTS-CLÉS : extraction d'information, événements médicaux, relations temporelles, médecine.

KEYWORDS: Information Extraction, Clinical Events, Temporal Relations, Medicine.

1 Introduction

La détection des relations temporelles entre événements dans un texte fournit des informations précieuses pour l'extraction d'information, la recherche de réponses à des questions, voire la traduction automatique. Les défis TempEval (Verhagen *et al.*, 2010) ont abordé cette problématique en « domaine ouvert », en cherchant à détecter (dans TempEval2) cinq types de relations temporelles (BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER) ou l'absence de relation déterminée (VAGUE). Ces relations étaient à trouver dans quatre situations : entre un événement et une date ou un autre événement qu'il domine syntaxiquement dans une phrase, entre un événement et la date de création du document, ou entre les deux événements principaux de deux phrases consécutives.

Identifier les informations temporelles décrivant la chronologie du séjour hospitalier d'un patient devrait amener une amélioration de la qualité de la prise en charge des patients (Harkema *et al.*, 2005; Zhou *et al.*, 2006; Savova *et al.*, 2009). Une meilleure connaissance de l'enchaînement des problèmes médicaux, des antécédents, des traitements, des rendez-vous, des opérations, permet en effet d'aider les analyses et les décisions des médecins ou des systèmes de surveillance

automatique. Les tâches et corpus du défi i2b2/VA 2012 (Sun *et al.*, 2013) ont ainsi été créés dans la perspective d'évaluer les méthodes d'extraction de ce type d'information à partir de textes cliniques. Nous nous intéressons ici à la tâche de détection des relations temporelles de ce défi.

Cette tâche diffère de la tâche TempEval présentée ci-dessus de plusieurs façons. Premièrement, elle ne considère que les trois premières relations (BEFORE, AFTER, OVERLAP) et l'absence de relation (que nous noterons NIL), les autres ayant un trop faible accord interannotateur lors de leur annotation humaine. Deuxièmement, elle ne restreint pas les situations où l'on peut trouver ces relations : dépendance syntaxique ou pas, phrases consécutives ou pas, etc. Troisièmement, les documents analysés, des comptes rendus hospitaliers, ont une structure qui les apparente à la concaténation de deux documents présentés dans deux sections successives clairement marquées : l'histoire de la maladie (*History of Present Illness* – HPI), telle que notée à l'entrée dans l'hôpital, et qui a comme date de référence la date d'entrée dans l'hôpital (*Admission Date* – AD) ; et ce qui s'est passé pendant le séjour hospitalier (*Hospital Course* – HC), tel que noté lors de la sortie de l'hôpital, et qui a comme date de référence la date de sortie de l'hôpital (*Discharge Date* – DD). Un compte rendu typique est présenté partiellement à la figure 1. On y voit les quatre sections toujours présentes ainsi que l'annotation fournie sur les événements et les expressions temporelles, et finalement les relations à trouver. Il peut également arriver qu'une relation relie des événements d'une section (HC) à l'autre (HPI).

ADMISSION DATE : <TIMEX3 type="date" id="t1">10/17/95</TIMEX3>

DISCHARGE DATE : <TIMEX3 type="date" id="t2">10/20/95</TIMEX3>

HISTORY OF PRESENT ILLNESS :
This is a 73-year-old man with <EVENT type="problem" id="e1">squamous cell carcinoma of the lung</EVENT>, status post <EVENT type="treatment" id="e2">lobectomy</EVENT> and <EVENT type="treatment" id="e3">resection</EVENT> of <EVENT type="problem" id="e4">left cervical recurrence</EVENT>, <EVENT type="occurrence" id="e5">admitted</EVENT> here with <EVENT type="problem" id="e6">fever</EVENT> and <EVENT type="problem" id="e7">neutropenia</EVENT>.
(...)

HOSPITAL COURSE :
He was started on <EVENT type="treatment" id="e8">Neupogen</EVENT>, 400 mcg. subq. q.d.
(...)
He was <EVENT type="occurrence" id="e9">discharged</EVENT> home on <EVENT type="treatment" id="e10">Neupogen</EVENT>.

```
e3 OVERLAP e4      e1 OVERLAP t1      e2 BEFORE t1      e3 BEFORE t1
e4 BEFORE t1       e6 OVERLAP t1      e7 OVERLAP t1     e8 BEFORE t2
e9 OVERLAP t2      e10 OVERLAP t2
```

FIGURE 1 – Extrait d'article (anonymisé) du corpus i2b2/VA, montrant les événements (EVENT), les expressions temporelles (TIMEX3) ainsi que quelques relations temporelles. On remarque que la notion d'événement est différente de ce qu'elle est souvent dans le domaine général : par exemple, des verbes d'action peuvent ne pas être des événements mais des noms de médicaments sont des événements, car ils désignent des traitements.

Ces deux dernières différences créent un nombre bien plus grand de situations où l'on peut rencontrer une relation temporelle que les quatre définies dans TempEval. La plupart des participants au défi se sont cependant focalisés sur quelques types de situations : relations à l'intérieur d'une même phrase, relations entre un événement et la date d'entrée ou de sortie, relations entre événements co-référents.

À notre connaissance, la formalisation de chaque situation et l'importance de ces situations sur le processus de détection des relations temporelles n'ont pas jusqu'ici été étudiées de manière systématique. Nous avons poussé cette logique à l'extrême en segmentant l'espace de détection des liens temporels en 57 situations distinctes. Nous avons étudié l'importance de ces situations sur notre corpus d'apprentissage en utilisant un oracle et avons déterminé empiriquement le classifieur qui correspondait le mieux à chaque situation.

Cet article étend celui présenté à l'atelier de clôture du défi i2b2/VA 2012 (Grouin *et al.*, 2012) par l'étude de ces situations oracle, réalisée depuis, et par les résultats complémentaires ainsi obtenus. Pour des raisons d'espace, nous nous y focalisons sur la découverte des relations temporelles entre des événements ou dates supposés déjà identifiés (nous décrivons la détection de ces événements et dates dans (Grouin *et al.*, 2012)). Nous présentons dans la suite notre méthode d'identification et d'étiquetage des relations temporelles (section 2) puis l'évaluation de la pertinence des différentes variantes proposées (section 3).

2 Identification des relations temporelles

Comme indiqué ci-dessus, la tâche consiste en premier lieu à choisir les paires sujettes à une relation, puis à typer celle-ci. Les quatre classes existantes sont donc BEFORE, AFTER, OVERLAP et NIL, la dernière signifiant « aucun lien ». Par ailleurs, on note que des relations importantes sont celles liant les dates d'admission (AD) et de sortie (DD) aux événements des sections d'histoire de la maladie (HPI) et de séjour hospitalier (HC).

Situations. Pour deux événements ou expressions temporelles données (collectivement notés EVT) qui peuvent faire l'objet d'une relation temporelle, plusieurs situations émergent selon leur section d'appartenance (AD, HPI, HC, AD), leur type (Event ou Timex3), leur présence dans la même phrase, la même section, etc. Nous avons ainsi identifié 56 combinaisons (plus « OTHER » pour les cas restants) des dimensions suivantes (nous utilisons le terme EVT pour désigner indifféremment un EVENT ou Timex3) :

- Section des EVT source et cible : *AD, DD, HPI, HC* ;
- Types d'éléments des EVT source et cible : *Timex3* ou *Event* ;
- Distance entre EVT : *même phrase (SS), phrases adjacentes de la même section (S1), phrases distantes* ;
- Nombre de Timex3 entre deux événements : aucune (NTB) ou au moins une.

Pour chaque combinaison, une méthode appropriée sera appliquée. Ainsi, *TIMEX3-EVENT-DD-HC* formalise une situation avec une expression temporelle (Timex3) dans la section « Discharge Date » *(la date de sortie elle-même)* et un événement (Event) dans la section « Hospital Course » (Figure 2). Ces distinctions sont importantes dans la mesure où, par exemple dans cette situation, la plupart des événements cités dans le séjour (HC) sont antérieurs à la date de sortie.

Pour toutes les paires d'EVT du corpus d'apprentissage correspondant à une situation, nous avons testé plusieurs classifieurs pour assigner à chaque paire l'une des quatre classes. Cette approche peut être comparée à un arbre de décision initial dont les caractéristiques seraient fondées sur ces quatre dimensions et pour lequel un autre classifieur serait appliqué à chaque feuille. Ainsi, une décision par défaut pour *TIMEX3-EVENT-DD-HC* pourrait être que l'événement intervient avant la date de sortie.

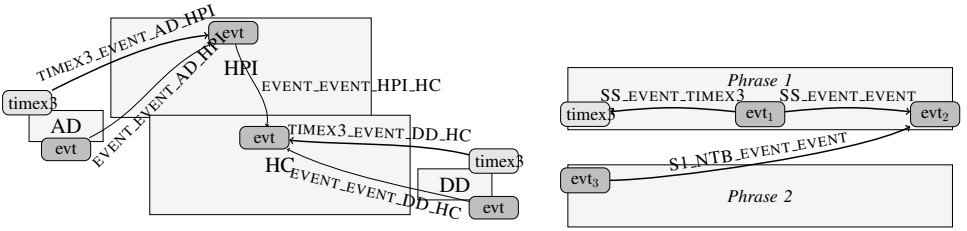

FIGURE 2 – Exemples de situations, liées à la période de la section (gauche : *_AD_HPI, *_DD_HC), dans la phrase courante ou dans la suivante (droite, SS*, S1*)

La question qui se pose alors est d'identifier les situations les plus importantes à traiter et les méthodes à utiliser. À notre connaissance cette question n'a pas été traitée jusqu'ici, y compris par les participants au défi i2b2/VA 2012 — et dans TempEval les situations étaient imposées. Nous avons pour cela calculé un rappel « oracle » pour chaque situation, c'est-à-dire la proportion de relations dans l'ensemble du corpus de référence correspondant à chaque situation. Ceci constitue le rappel idéal qui devrait être atteint par un classifieur entraîné pour cette situation. Des expériences nous ont montré qu'il est impossible d'obtenir un gain pour les situations avec un rappel oracle inférieur à 1 % ; nous nous sommes donc focalisés sur les situations avec rappel oracle supérieur à ce seuil.

Caractéristiques d'apprentissage. Les caractéristiques choisies pour l'apprentissage sont à la fois internes et contextuelles :

- Caractéristiques internes :
  - Toutes les annotations de l'EVT (modalité, polarité, sous-types) ;
  - Pour les événements, une sous-catégorisation d'après une étude du lexique dans le jeu d'apprentissage *(chirurgie, état, événement ponctuel, localisation, suivi d'événement, autres)* ;
  - Le texte de l'EVT (si parmi les 50 les plus fréquents) ;
  - Les mots de l'EVT (si présents au moins 10 fois dans le corpus d'apprentissage) ;
  - Nous avons également testé des classes distributionnelles (clusters de Brown) sur les mots des EVT, combinées aux prépositions temporelles, mais cela n'a permis aucune amélioration.
- Caractéristiques contextuelles :
  - La distance entre les deux EVT de la relation ;
  - La présence de prépositions temporelles ou non entre EVT ;
  - Le nombre d'autres EVT entre les deux ;
  - Pour les paires d'événements dans une même phrase, les dépendances syntaxiques obtenues par l'analyseur syntaxique Charniak McClosky converties en dépendances de Stanford ;
  - Des combinaisons de ces traits, par exemple *<type-EVT-source, dépendance-syntaxique, type-EVT-cible>*, où les types associés à *source* et *cible* sont les catégories i2b2 *(clinical department, evidential, occurrence, problem, test, treatment ; date, duration, frequency, time)*.

Absence de relation et fermeture transitive. La définition de la tâche implique que les annotations manuelles sont incomplètes : certaines instances positives des relations temporelles peuvent être inférées par une fermeture transitive depuis les annotations fournies par la référence, et ne doivent donc pas être considérées comme des exemples négatifs. Lors de l'apprentissage, nous avons donc appliqué, comme Mani *et al.* (2006), une fermeture transitive de toutes les relations

temporelles avec l'outil Sputlink (Verhagen et Pustejovsky, 2008), les liens produits étant alors utilisés comme des instances positives. Le produit croisé des EVT de chaque document a servi à générer un ensemble complet de relations temporelles candidates, duquel les instances positives ont été retirées de manière à générer un ensemble d'instances négatives. La fermeture transitive n'a en revanche pas été appliquée lors de l'application sur les corpus d'apprentissage et de test des modèles ainsi obtenus.

Différents classifieurs. Les résultats de Costa et Branco (2013) montrent que les meilleurs résultats dans chaque situation sont obtenus par différents classifieurs (dans leur cas, tables de décision, arbre de décision, JRip, K étoile, classifieur bayésien naïf). Le classifieur à arbres de décision J48, implémenté dans Weka à partir de l'algorithme C4.5, a obtenu dans notre cas les meilleurs résultats. Nous avons également évalué les résultats obtenus par d'autres classifieurs (bayésien naïf, séparateur à vaste marge (LibSVM), k plus proches voisins, régression logistique (MaxEnt), forêt d'abres décisionnels) pour les principales situations [1]. Nous avons retenu le meilleur classifieur pour chaque situation et appliqué ce classifieur dans chaque instance de la situation correspondante. Nous avons également testé une combinaison des classifieurs par un vote, en utilisant la moyenne de leurs confiances respectives comme opérateur de combinaison.

Sachant qu'une relation temporelle prédite par un classifieur (ou inférée par la fermeture transitive) peut entrer en contradiction avec une relation précédemment prédite, nous avons traité les situations par ordre décroissant de leur précision sur le jeu d'apprentissage. En cas d'incohérence, de manière similaire à (Mani *et al.*, 2007), la nouvelle relation prédite est écartée.

Décisions à base de règles. Certaines relations peuvent être identifiées à partir de simples règles : l'admission est avant la sortie, la date et les événements liés à l'admission se chevauchent.

3 Évaluation et discussion

Le corpus d'apprentissage comprend 190 comptes rendus contre 120 dans le corpus de test. Les objectifs du défi (Sun *et al.*, 2013) concernent l'identification de six types d'événements *(département clinique, preuve, occurrence, problème, examen, traitement)*, les expressions temporelles *(Timex3 : date, durée, fréquence, heure)* et les relations temporelles (*Tlinks* : de type BEFORE, OVERLAP ou AFTER). Nous ne considérons que les relations temporelles dans cet article. Les mesures d'évaluation utilisées sont décrites par Sun *et al.* (2013).

La figure 3 (gauche) [2] représente le rappel oracle (*RO*) dans chaque situation du corpus d'apprentissage, par ordre décroissant, sur une échelle semi-logarithmique. Les situations les plus contributives ($RO > 0,01$) concernent trois situations intra-phrase (SS-*, somme $RO = 0,48$), quatre situations en lien avec les dates d'admission ou de sortie (*-AD-HPI, *-DD-HC, somme $RO = 0,31$), une situation entre deux phrases successives (S1-NTB-EVENT-EVENT, $RO = 0,04$), la co-référence (SAME-TEXT, $RO = 0,02$), et une relation à plus longue distance entre événements durant le séjour hospitalier (NTB-EVENT-EVENT-HC-HC, $RO = 0,01$). Nous nous sommes intéressés aux huit premières, abandonnant ainsi un rappel de 0,17 (dont les situations « autres », $RO = 0,07$). Nous avons par ailleurs constaté que la situation EVENT-TIMEX3-AD-HPI pouvait

1. Nous avons fait face à une limitation technique dans l'usage de la régression logistique de Weka, dont le nombre d'attributs de type « mots de l'EVT » a dépassé les capacités. Nous l'avons donc utilisée sans cet attribut.

2. La lecture de cette figure sur impression papier donne une idée générale de la distribution observée, son examen détaillé est possible lors de sa lecture à l'écran, où un facteur de grandissement peut être appliqué.

obtenir une bonne précision et l'avons ajoutée à l'ensemble des huit situations conservées.

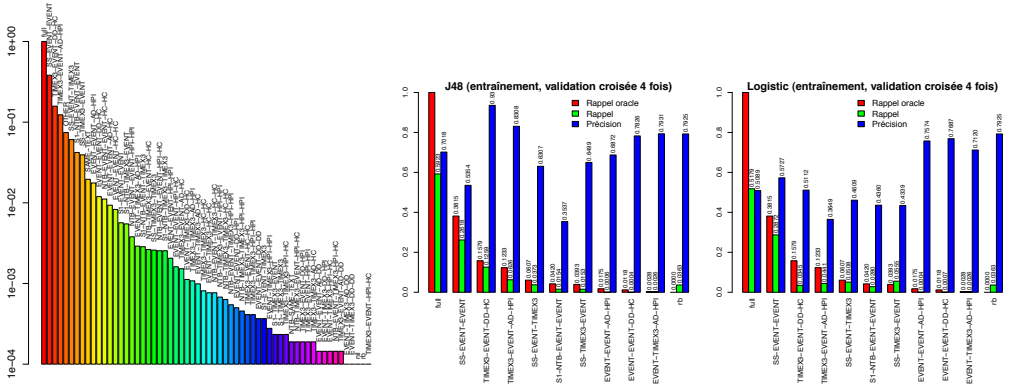

FIGURE 3 – Rappel oracle dans chaque situation du corpus d'apprentissage par ordre décroissant (gauche), performance de J48 sur les 9 meilleures situations (centre), performance de la régression logistique sur les 9 meilleures situations (droite).

La figure 3 montre également les performances des classifieurs par arbres de décision J48 (centre) et par régression logistique (droite) dans chacune de ces neuf situations, auxquelles nous avons ajouté les décisions à base de règles *(rb, rule-based)* et le total *(full)*. La précision de J48 est généralement bonne (moindre dans une même phrase). Grâce au rappel oracle, nous constatons que son rappel est bon sur TIMEX3-EVENT-DD-HC et modéré sur SS-EVENT-EVENT et TIMEX3-EVENT-AD-HPI. La distance avec le rappel oracle étant importante dans ces deux cas (resp. 10 points pour Logistique et 6 points pour J48), une étude complémentaire est indiquée dans ces deux situations. La régression logistique obtient une précision plus basse et un meilleur rappel que J48 pour les situations dans une même phrase, mais est moins bon sur les liens entre date d'admission ou de sortie (AD ou DD) et les événements des autres sections (HPI ou HC).

Une interprétation de la bonne performance des arbres de décision est qu'ils ont la capacité de construire des conjonctions de caractéristiques, utiles pour ces situations, alors que la régression logistique ne permet pas de le faire (elle permet seulement d'ajouter les scores obtenus individuellement par chaque caractéristique, mais pas de produire d'elle-même une conjonction de caractéristiques qui obtiendrait un score plus important que cette somme). La construction a priori de telles caractéristiques combinées a augmenté les scores de ces autres classifieurs. Nous avons également testé un séparateur à vaste marge (LibSVM), mais il n'a pas montré de performances compétitives dans les situations testées. Les forêts d'arbres aléatoires semblent plus robustes pour les situations avec un faible nombre d'instances.

Le tableau 1 contient les résultats obtenus par l'étude systématique de ces situations, examinant initialement 20 situations (avant étude du rappel oracle) puis les 9 plus productives. Toutes nos études ont été réalisées sur le corpus d'apprentissage avec une validation croisée (mais seulement en quatre parties du fait du temps élevé d'apprentissage). Grâce à ces précautions, toutes les améliorations réalisées lors de l'optimisation sur le corpus d'apprentissage ont également été reportées sur le corpus de test.

Utiliser les mots de l'EVT comme caractéristiques (mention « mots » dans le tableau) a géné-

		Apprentissage (v-c 4 fois)			Corpus de test			
#S	Classifieur	P	R	F	P	R	F	Delta
9	Sel[P] (mots)	0.711	0.628	0.667	**0.655**	0.594	**0.623**	-0.044
9	C	0.652	0.663	0.658	0.602	0.631	0.616	-0.042
9	j48	0.702	0.592	0.642	0.661	0.555	0.603	-0.039
9	rf (mots)	0.661	0.591	0.624	0.628	0.557	0.590	-0.034
9	rf	0.624	0.573	0.598	0.580	0.549	0.564	-0.034
9	nb (mots)	0.569	0.581	0.575	0.390	0.520	0.445	-0.129
9	logistic	0.509	0.518	0.513	0.345	0.470	0.398	-0.115
9	nb	0.478	0.537	0.506	0.331	0.454	0.383	-0.123
20	Sel[F]				0.601	**0.615**	**0.608**	
20	vote				0.654	0.549	0.597	
20	j48				0.644	0.532	0.583	
20	LibSVM				**0.659**	0.511	0.576	
20	rf				0.589	0.522	0.553	

Légende : v-c=validation croisée ; #S=nombre de situations, P=précision, R=rappel, F=F-mesure, Delta=différence en F-mesure entre le corpus de test et celui d'apprentissage. Sel=Sélection des différents classifieurs pour chaque section, en optimisant la précision [P] ou la F-mesure [F]. rf=forêt aléatoire, nb=bayésien naïf. Le vote (moyenne des prédictions) combine J48, LibSVM, forêt aléatoire, k plus proches voisins, et bayésien naïf. Mots=incluant les mots de l'EVT comme caractéristiques. Les 9 situations sont ordonnées par précision décroissante, les 20 situations sont ordonnées approximativement. Les cellules vides renvoient à des données non disponibles.

TABLE 1 – Résultats globaux sur les relations temporelles.

ralement amélioré les résultats des classifieurs (sauf pour J48) ; l'impossibilité d'utiliser ces caractéristiques pour la régression logistique dans Weka en a certainement limité les performances. Nous comptons donc tester un autre classifieur à maximum d'entropie.

Nous avons commencé l'étude des situations qui ont un rappel oracle plus faible, mais jusqu'ici les augmentations de rappel n'ont pas compensé les pertes de précision associées. La recherche de nouvelles caractéristiques constitue une piste complémentaire à explorer.

4 Conclusion

Dans cet article, nous avons montré comment une étude systématique des situations où l'on peut rencontrer des relations temporelles dans des comptes rendus hospitaliers, incluant le calcul du rappel oracle de ces situations et une comparaison de différents classifieurs, nous a permis d'obtenir des résultats sensiblement meilleurs que ceux obtenus sans effectuer cette étude.

Plusieurs pistes d'amélioration existent. En particulier, formaliser des caractéristiques supplémentaires dédiées à chaque situation et utiliser des implémentations de classifieurs qui passent à l'échelle devraient améliorer nos performances. Au lieu d'une procédure de décision gloutonne, une procédure de décision globale pourrait être implémentée pour étudier le graphe de toutes les relations temporelles prédites, y compris les relations en conflit, en tenant compte de la confiance de leur prédiction, et devrait permettre la sélection d'un sous-graphe cohérent avec un score de prédiction global optimal. En ce sens, Costa et Branco (2013) proposent d'utiliser les

informations produites dans les situations déjà traitées pour prédire les relations temporelles dans le reste des situations à traiter. Enfin, la caractérisation du rappel oracle nous a permis de mettre en évidence les directions à améliorer : les relations à l'intérieur d'une phrase et la mise en relation avec la date d'admission des événements de l'histoire de la maladie.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme Quaero, financement Oseo (agence nationale de valorisation de la recherche) et du projet Accordys (ANR-12-CORD-0007-03). Les données médicales proviennent du consortium Informatics for Integrating Biology to the Bedside (i2b2) grâce aux financements 2U54LM008748 du NIH/National Library of Medicine (NLM), National Heart, Lung and Blood Institute (NHLBI), et 1R13LM01141101 du NIH/NLM.

Références

COSTA, F. et BRANCO, A. (2013). Temporal relation classification based on temporal reasoning. *In Proc International Workshop on Computational Semantics*, Potsdam, Allemagne. ACL SIGSEM.

GROUIN, C., GRABAR, N., HAMON, T., ROSSET, S., TANNIER, X. et ZWEIGENBAUM, P. (2012). A tale of temporal relations between clinical concepts and temporal expressions : towards a representation of the clinical patient's timeline. *In* UZUNER, O., SUN, W. et RUMSHISKY, A., éditeurs : *i2b2/VA Workshop Proc*, Chicago, IL. i2b2. 9 pages.

HARKEMA, H., SETZER, A., GAIZAUSKAS, R. et HEPPLE, M. (2005). Mining and modelling temporal clinical data. *In The UK e-Science All Hands Meeting Proc*, pages 507–514.

MANI, I., VERHAGEN, M., WELLNER, B., LEE, C. M. et PUSTEJOVSKY, J. (2006). Machine learning of temporal relations. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics.

MANI, I., WELLNER, B., VERHAGEN, M. et PUSTEJOVSKY, J. (2007). Three approaches to learning TLINKs in TimeML. Technical Report CS-07-268, Brandeis University.

SAVOVA, G., BETHARD, S., STYLER, W., MARTIN, J., PALMER, M., MASANZ, J. et WARD, W. (2009). Towards temporal relation discovery from the clinical narrative. *In AMIA Annu Symp Proc*, pages 568–572.

SUN, W., RUMSHISKY, A. et UZUNER, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge overview. *J Am Med Inform Assoc*. Soumis.

VERHAGEN, M. et PUSTEJOVSKY, J. (2008). Temporal processing with the TARSQI toolkit. *In Coling Proc*, pages 189–192. Démonstration.

VERHAGEN, M., SAURI, R., CASELLI, T. et PUSTEJOVSKY, J. (2010). Semeval-2010 task 13 : Tempeval-2. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

ZHOU, L., MELTON, G., PARSONS, S. et HRIPCSAK, G. (2006). A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform*, 39(4):424–439.

