Extraction de relations temporelles dans des dossiers électroniques patient

Julien Tourille[1] Olivier Ferret[2] Aurélie Névéol[3] Xavier Tannier[1]
(1) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay
(2) CEA, LIST, Gif-sur-Yvette, F-91191 France
(3) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
prenom.nom@limsi.fr, prenom.nom@cea.fr

RÉSUMÉ

L'analyse temporelle des documents cliniques permet d'obtenir des représentations riches des informations contenues dans les dossiers électroniques patient. Cette analyse repose sur l'extraction d'événements, d'expressions temporelles et des relations entre eux. Dans ce travail, nous considérons que nous disposons des événements et des expressions temporelles pertinents et nous nous intéressons aux relations temporelles entre deux événements ou entre un événement et une expression temporelle. Nous présentons des modèles de classification supervisée pour l'extraction de des relations en français et en anglais. Les performances obtenues sont comparables dans les deux langues, suggérant ainsi que différents domaines cliniques et différentes langues pourraient être abordés de manière similaire.

ABSTRACT

Extracting Temporal Relations from Electronic Health Records.

Temporal analysis of clinical documents yields complex representations of the information contained in Electronic Health Records. This type of analysis relies on the extraction of medical events, temporal expressions and the relations between them. In this work, we assume that relevant events and temporal expressions are available and we focus on the extraction of relations between two events or between an event and a temporal expression. We present supervised classification models and apply them to clinical documents written in French and in English. The performance we achieve is high and similar for both languages. We believe these results suggest that temporal analysis may be approached generically across clinical domains and languages.

MOTS-CLÉS : Extraction de Relations, Analyse Temporelle, Traitement de la Langue Biomédicale.

KEYWORDS: Relation Extraction, Temporal Analysis, Medical Language Processing.

1 Introduction

La détection automatique des événements et des relations temporelles entre ces événements dans les documents cliniques est une tâche récente à laquelle de nombreuses équipes s'intéressent pour la langue anglaise. Les campagnes d'évaluation i2b2 (Sun *et al.*, 2013) et Clinical TempEval (Bethard *et al.*, 2015, 2016) ont permis de donner un cadre à ces efforts. Les documents cliniques se distinguent des textes généralistes et d'autres genres de textes bio-médicaux par le fait que chaque dossier patient, contenant des comptes-rendus d'actes, des comptes-rendus de séjour et des courriers, relate l'histoire médicale et le parcours de soin d'un patient. Il serait alors très intéressant pour l'équipe soignante

d'avoir accès à des outils permettant de résumer cette histoire (Hirsch *et al.*, 2015) ou de la comparer à un protocole de traitement standard.

Dans cet article, nous nous intéressons particulièrement à la détection des relations d'inclusion temporelle (X contient Y) intra-phrastiques entre les événements et/ou expressions temporelles (CR pour *Container Relation*) et aux relations temporelles entre les événements et la date de création du document (DR pour *Doctime Relation*). La relation d'inclusion temporelle est directement liée à celle de conteneur narratif et a été introduite par Pustejovsky & Stubbs (2011). L'utilisation de conteneurs narratifs permet de réduire les inconsistances dans les annotations et offre une plus grande précision de modélisation des relations temporelles. Le corpus THYME (Styler IV *et al.*, 2014), utilisé jusqu'à présent dans les campagnes Clinical TempEval est le premier corpus clinique annoté avec le concept de conteneur narratif. La détection de ces relations est une première étape vers la création d'une chronologie détaillée des événements considérés. Une perspective de recherche serait d'utiliser ces relations pour la création de graphes temporels (Bramsen *et al.*, 2006). Dans cet article, nous considérons que les événements ont déjà été extraits. Nous décrivons des expérimentations sur deux corpus, l'un en anglais et l'autre en français.

Cet article présente et compare tout d'abord les corpus utilisés dans les deux langues ainsi que la définition de la notion d'événement couramment employée dans le domaine clinique. Nous décrivons ensuite une approche d'extraction des relations en précisant les différences de traitement selon la langue considérée. Nous présentons et discutons enfin les résultats obtenus.

2 Présentation des corpus

Pour le français, nous avons utilisé des documents d'un corpus de textes cliniques issus d'un groupe d'institutions hospitalières françaises. Pour ce travail, nous avons sélectionné des documents issus du service d'hépato-gastro-nutrition ayant déjà fait l'objet d'une désidentification (Grouin & Névéol, 2014) et d'un travail d'annotation en entités et relations (Deléger *et al.*, 2014). Pour l'anglais, nous avons utilisé le corpus mis à disposition pour la campagne Clinical TempEval 2016 (Bethard *et al.*, 2016), comportant des documents issus du service de cancérologie de la Mayo Clinic aux États-Unis. Le tableau 1 présente les caractéristiques de ces corpus.

La définition d'un événement varie selon le corpus. Le corpus Clinical TempEval annote comme événement tout ce qui présente un intérêt du point de vue de la chronologie de prise en charge d'un patient comme par exemple une maladie ou une procédure médicale. Aucune information sémantique n'est donnée concernant ces événements. Seuls les empans figurent dans les annotations. Dans le corpus français, outre ces empans, une définition des événements est formalisée à l'aide des catégories sémantiques de l'UMLS (Unified Medical Language System) : *disorder*, *sign or symptom*, *medical procedure*, *chemical drugs*, *concept or idea*, *biological process or function* (Deléger *et al.*, 2014).

3 Méthodologie

Nous avons abordé les deux tâches comme des problèmes de classification supervisée. Pour la tâche DR, chaque événement mentionné dans le dossier se voit assigner une catégorie parmi les quatre suivantes : *Before*, *Before-Overlap*, *Overlap* et *After*. La tâche CR est quant à elle un problème de classification binaire appliquée à chaque paire d'événement et/ou expression temporelle. Cependant,

	Corpus TempEval	Corpus français
Nombre de tokens	463 091	167 013
Nombre d'événements[a]	DR 59 976 CR 49 147	15 469
Nombre d'expressions temporelles	6 191	3 365
Nombre de relations intra-phrastiques	13 319	3 642

[a] Certains documents du corpus TempEval ne sont pas annotés en relations. Nous distinguons donc le nombre d'événements utiles dans la tâche DR de ceux de la tâche CR.

TABLE 1 – Présentation des corpus : statistiques descriptives

considérer toutes les paires d'entités au sein des documents entraînerait la construction d'un ensemble d'apprentissage déséquilibré où le nombre d'exemples positifs (présence d'une relation) serait très largement inférieur au nombre d'exemples négatifs (absence de relation). Afin de réduire le nombre de paires candidates au sein des ensembles d'entraînement et de test, nous avons transformé le problème de classification à deux classes (*contient vs. pas-de-relation*) en un problème à trois classes (*contient*, *est-contenu* et *pas-de-relation*). Au lieu de considérer toutes les permutations d'entités au sein d'une phrase, nous considérons donc toutes les combinaisons d'événements de gauche à droite, en changeant le sens et l'étiquette des relations (*contient* $\rightarrow$ *est-contenu*) quand cela s'avère nécessaire. Cette transformation nous a permis de diviser par deux le nombre de paires candidates dans les corpus, avec 111 447 paires pour le corpus anglais et 67 113 paires pour le corpus français. Lors de la phase d'évaluation, les prédictions *est-contenu* sont changées en prédictions *contient*.

Par ailleurs, certaines entités sont plus susceptibles que d'autres de contenir des événements. C'est le cas des expressions temporelles qui sont, de par leur nature, des candidats préférentiels. C'est aussi le cas de certains événements médicaux que l'on peut qualifier de complexe. L'événement *opération chirurgicale/surgical operation* peut ainsi contenir d'autres événements, comme *suture/suturing* ou *saignement/bleeding*, alors que ce n'est généralement pas le cas de ces deux derniers, considérés comme élémentaires. En suivant cette observation, nous avons construit un classifieur binaire d'événement et d'expressions temporelles ayant pour but de déterminer si l'entité considérée est un conteneur potentiel. Le résultat de ce classifieur est ensuite utilisé comme attribut dans le classifieur de relations.

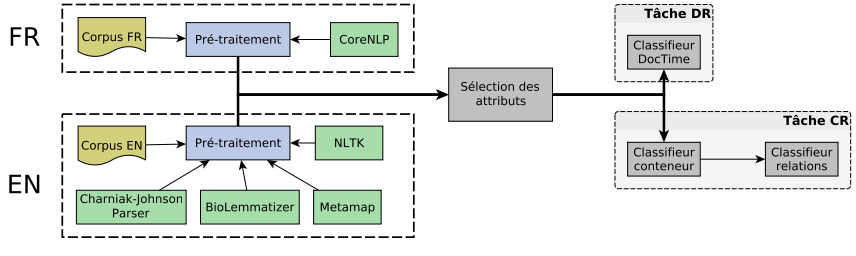

FIGURE 1 – Vue d'ensemble de la chaîne de traitement

La figure 1 donne une vue synthétique des différents processus mis en œuvre pour traiter ces deux tâches. Le pré-traitement des corpus est détaillé à la section 4.3. Pour les différents classifieurs mis en place dans les tâches DR et CR, nous avons extrait des attributs structurels, contextuels et lexicaux. Ces attributs sont présentés au tableau 2. Les tailles optimales des fenêtres pour les contextes gauche et droit ont été calculées par validation croisée sur le corpus d'entraînement.

Attributs	DocTime	Conteneur	Relation
Type des entités	✓	✓	✓
Formes des entités	✓	✓	✓
Attributs des entités[a]	✓	✓	✓
Positions relatives des entités dans le document	✓	✓	✓
Entités classées comme conteneurs potentiels			✓
Type du document[b]	✓	✓	✓
Formes des entités dans les contextes gauche, centre et droit[c]	✓	✓	✓
Types des entités dans les contextes gauche, centre et droit[c]	✓	✓	✓
Attributs des entités dans les contextes gauche, centre et droit[ac]	✓	✓	✓
Entités marquées comme conteneur dans les contextes gauche, centre et droit			✓
Étiquettes POS des verbes de la phrase	✓	✓	
Formes des tokens contextuels (unigrammes)	✓	✓	
Étiquettes morpho-syntaxiques des tokens contextuels (unigrammes)	✓	✓	
Formes des tokens contextuels (bigrammes)[d]	✓	✓	
Étiquettes POS des tokens contextuels (bigrammes)[d]	✓	✓	

[a] Dans le corpus anglais, plusieurs attributs sont disponibles pour les événements : la *modalité* (actual, hypothetical, hedged ou generic), le *degré* (most, little), la *polarité* (pos ou neg) et le *type* (aspectual, evidential ou N/A). Pour les expressions temporelles, l'attribut *classe* (date, time, duration, quantifier, prepostexp ou set) concerne les deux langues, l'attribut *type* (date, duration, frequency ou time), le seul français.
[b] Cette information est disponible seulement pour le corpus français.
[c] Le contexte centre est considéré seulement pour le modèle *Relation*.
[d] Seulement dans le cas où l'on utilise les formes fléchies des tokens.

TABLE 2 – Attributs utilisés par les classifieurs de notre chaîne de traitement

4 Expérimentations

4.1 Représentation des attributs lexicaux

Nous avons implémenté deux stratégies pour représenter les attributs lexicaux dans nos classifieurs. Dans la première, nous avons utilisé les formes fléchies des tokens telles qu'elles figurent dans les textes. Dans la seconde stratégie, nous avons remplacé ces attributs par leur représentation vectorielle calculée avec l'outil *word2vec* (Mikolov *et al.*, 2013) sur des corpus cliniques bruts. Dans le cas du français, nous avons utilisé l'intégralité des textes disponibles, non annotés, et désidentifiés selon le même protocole que le corpus annoté. La taille finale du corpus est d'environ 37 millions de mots. En ce qui concerne l'anglais, nous avons utilisé le corpus Mimic 2 (Saeed *et al.*, 2011) pour calculer les vecteurs [1]. La taille du corpus est d'environ 17 millions de mots.

Dans un certain nombre de cas, nos attributs sont des unités polylexicales. Pour construire leur représentation vectorielle, nous avons adopté une méthode classique de *max pooling* consistant à prendre la valeur maximale de chaque dimension parmi les vecteurs des tokens constituant ces unités polylexicales. Chaque contexte et chaque entité considérée, événement ou expression temporelle, sont donc représentés par un vecteur de 200 dimensions.

1. Paramètres utilisés : -min-count 5 -size 200 -window 20 -sample 1e-3 -cbow 1 pour l'anglais, -min-count 5 -size 200 -window 10 -sample 1e-3 -cbow 1 pour le français

4.2 Sélection des algorithmes

Les corpus ont été divisés en ensembles d'apprentissage et de test selon le ratio 80/20. Nous avons réalisé une recherche de type *grid search* pour sélectionner l'algorithme approprié ainsi que ses paramètres à chaque étape de notre chaîne de traitement. Dans les trois cas et pour les deux stratégies, nous avons considéré deux algorithmes de classification supervisée : forêt d'arbres décisionnels (*Random Forests*) et machine à vecteurs de support (*LinearSVM*, avec l'implémentation *liblinear*).

Dans chaque cas, une validation croisée à 5 plis a été effectuée pour choisir l'algorithme et ses paramètres. Dans le cas des modèles *Conteneur* et *DocTime*, nous avons utilisé l'exactitude (*accuracy*) comme mesure d'évaluation. En ce qui concerne le modèle *Relation*, nous avons utilisé la F1-mesure. Les algorithmes et les paramètres retenus par la validation croisée sont présentés au tableau 3. L'implémentation de ces modèles a été réalisée avec la bibliothèque Python d'apprentissage Scikit-learn (Pedregosa *et al.*, 2011).

Corpus	Modèle	Algorithme	Paramètres
FR	DocTime	Random Forests	fenêtre=3, nbmax. attributs=auto, nb. arbres=50, mesure=entropy, w2v=oui
	Conteneur	LinearSVM	fonction objectif=square_hinge, régularisation=l2, C=1, fenêtre=6, tol=0,01, w2v=non
	Relation	LinearSVM	fonction objectif=hinge, régularisation=l2, C=1, tol=0,001, w2v=oui
EN	DocTime	LinearSVM	fonction objectif=hinge, régularisation=l2, C=1, fenêtre=6, tol=0,0001, w2v=oui
	Conteneur	Random Forests	fenêtre=1, nbmax. attributs=sqrt, nb. arbres=100, mesure=gini, w2v=non
	Relation	LinearSVM	fonction objectif=hinge, régularisation=l2, C=1, tol=0,01, w2v=oui

TABLE 3 – Algorithmes d'apprentissage et paramètres utilisés en fonction du corpus, pour les classifieurs de la chaîne de traitement

4.3 Pré-traitement des corpus de travail

Contrairement à l'anglais, il existe peu de ressources disponibles pour le traitement automatique de la langue biomédicale en français. De ce fait, nous avons dû utiliser des ressources adaptées au domaine général. La première étape consiste à segmenter, tokeniser et étiqueter en parties du discours. Pour l'anglais, nous avons utilisé NLTK pour la segmentation (Loper & Bird, 2002) et le BLLIP Reranking Parser (Charniak & Johnson, 2005) en association avec un modèle pré-entraîné sur un corpus biomédical (McClosky, 2010) pour la tokenisation et l'analyse morpho-syntaxique. Pour le français, nous avons utilisé CoreNLP (Manning *et al.*, 2014) et le modèle pré-entraîné sur le français pour le domaine général pour la segmentation, la tokenisation et l'analyse morpho-syntaxique.

Enfin, concernant la catégorisation sémantique des événements dans les textes, nous avons adopté des stratégies différentes selon les deux corpus considérés : dans le cas du corpus français, nous nous sommes appuyés sur la catégorisation déjà existante des entités mentionnée dans la section 2 tandis que pour le corpus anglais, nous avons exploité les résultats de l'outil Metamap (Aronson & Lang, 2010) pour déterminer les types des événements présents dans le corpus.

5 Résultats et discussion

Les résultats de la validation croisée pour les différents classifieurs ainsi que les résultats obtenus sur l'ensemble de test sont présentés dans le tableau 4.

		DocTime		Conteneur		Relation		Relation sans Conteneur	
Corpus	Algorithme	Normal	W2V	Normal	W2V	Normal	W2V	Normal	W2V
FR	Random Forests	0,752 (0,008)	0,775 (0,007)	0,940 (0,004)	0,942 (0,006)	0,596 (0,009)	0,686 (0,015)	0,382 (0,014)	0,574 (0,009)
	Linear SVM	0,770 (0,010)	0,766 (0,009)	0,945 (0,003)	0,928 (0,006)	0,749 (0,013)	0,749 (0,016)	0,555 (0,021)	0,569 (0,020)
EN	Random Forests	0,820 (0,002)	0,803 (0,002)	0,925 (0,003)	0,922 (0,004)	0,657 (0,012)	0,571 (0,006)	0,510 (0,012)	0,538 (0,012)
	Linear SVM	0,832 (0,010)	0,842 (0,004)	0,914 (0,005)	0,916 (0,004)	0,735 (0,006)	0,748 (0,004)	0,579 (0,011)	0,588 (0,007)

(a) Résultats des différents modèles par validation croisée sur le corpus d'entraînement. Nous reportons l'exactitude (accuracy) pour les modèles *DocTime* et *Conteneur* et la F1-mesure pour le modèle *Relation*. Pour chaque mesure, nous précisons l'écart type (entre parenthèses).

	Corpus FR			Corpus EN		
Relation	P	R	F1	P	R	F1
Bef./Over.	0,65	0,48	0,55	0,68	0,53	0,59
Before	0,85	0,44	0,58	0,87	0,83	0,85
After	0,81	0,43	0,56	0,81	0,82	0,81
Overlap	0,79	0,94	0,86	0,85	0,90	0,87
Moyenne	0,78	0,78	0,76	0,84	0,84	0,84

(b) Résultats obtenus sur le corpus de test par le modèle *DocTime*. Nous reportons la précision (P), le rappel (R) et la F1-mesure (F1) pour chaque type de relation en fonction du corpus utilisé.

	Corpus FR			Corpus EN		
	P	R	F1	P	R	F1
pas de rel.	0,98	0,98	0,98	0,94	0,96	0,95
contient	0,62	0,59	0,61	0,62	0,48	0,54
Moyenne	0,96	0,96	0,96	0,90	0,91	0,91

(c) Résultats obtenus sur le corpus de test par le modèle *Relation*. Nous reportons la précision (P), le rappel (R) et la F1-mesure (F1) pour chaque type de relation en fonction du corpus utilisé.

TABLE 4 – Présentation des résultats obtenus lors des expérimentations

Nous obtenons des résultats satisfaisants dans les deux tâches CR et DR. Pour la tâche DR, on observe un certain écart entre les F-mesures observées pour l'anglais (0,84) et le français (0,76). On note par ailleurs que les résultats par catégorie ne sont pas homogènes pour les deux langues. Pour le français, on observe un écart d'environ 0,30 entre les F1-mesures des catégories *Before-Overlap*, *Before* et *After* d'un côté et *Overlap* de l'autre. En ce qui concerne l'anglais, le modèle ne parvient pas à obtenir de bons résultats dans la catégorie *Before-Overlap* (0,59 de F1-mesure) mais parvient à des résultats homogènes autour de 0,80 pour les autres catégories. En ce qui concerne la tâche CR, les résultats obtenus sont proches avec une F1-mesure de 0,61 pour le corpus français sur la catégorie *contient* et une F1-mesure de 0,54 pour l'anglais. On observe que la précision pour les deux langues est identique. Le rappel en revanche est en dessous de la moyenne pour l'anglais. A titre de comparaison, le meilleur score obtenu lors de l'édition 2016 de Clinical TempEval est 0,843 (accuracy) pour la tâche DR et 0,573 (F1-mesure) pour la tâche CR.

L'utilisation des word embeddings par rapport à l'utilisation des formes fléchies semble avoir un impact positif pour les classifieurs *DocTime* et *Relation* mais reste relativement limité. En ce qui

concerne le premier, le gain en termes de F1-mesure est faible (+0,005 pour le français et +0,01 pour l'anglais). En ce qui concerne le modèle *Relation*, le gain est nul pour le français et de +0,013 pour l'anglais. Le modèle *Conteneur* ne semble pas bénéficier de l'utilisation des word embeddings. On observe une perte de -0,003 en termes de F1-mesure pour le français et l'anglais.

En ce qui concerne les différences de performance observées entre les deux langues, plusieurs hypothèses peuvent être formulées. Tout d'abord, l'utilisation de ressources spécialisées pour le traitement de la langue biomédicale en français permettrait d'obtenir de meilleurs résultats quant au pré-traitement des textes et pourrait donc améliorer les résultats finaux. Ensuite, les corpus anglais et français sont déséquilibrés en termes de volume et d'entités annotées. Les performances sur le français peuvent ainsi être touchées par le faible nombre de textes annotés. Enfin la qualité des annotations, notamment celle des annotions en événements, qui est plus formelle et plus fine pour le corpus français que pour le corpus anglais, peut influencer les performances du système, notamment pour le modèle *Relation*.

6 Conclusion

Nous avons présenté un système d'extraction de relations temporelles dans des documents cliniques extraits de dossiers électroniques patient. Les expérimentations que nous avons menées sur des corpus anglais et français nous permettent d'obtenir des résultats satisfaisants et comparables pour les deux tâches d'extraction DR et CR et semblent donc indiquer que le traitement des relations temporelles dans des textes cliniques peut se généraliser au-delà de ces deux langues.

Remerciements

Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen pour la mise à disposition du corpus LERUDI. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche sous la référence CABeRneT ANR-13-JS02-0009-01.

Références

ARONSON A. R. & LANG F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, **17**(3), 229–236.

BETHARD S., DERCZYNSKI L., SAVOVA G., PUSTEJOVSKY J. & VERHAGEN M. (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 806–814, Denver, USA: Association for Computational Linguistics.

BETHARD S., SAVOVA G., CHEN W.-T., DERCZYNSKI L., PUSTEJOVSKY J. & VERHAGEN M. (2016). SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California: Association for Computational Linguistics.

BRAMSEN P., DESHPANDE P., LEE Y. K. & BARZILAY R. (2006). Inducing Temporal Graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*,

EMNLP '06, p. 189–198, Stroudsburg, PA, USA: Association for Computational Linguistics.

CHARNIAK E. & JOHNSON M. (2005). Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 173–180, Ann Arbor, Michigan: Association for Computational Linguistics.

DELÉGER L., GROUIN C., LIGOZAT A.-L., ZWEIGENBAUM P. & NÉVÉOL A. (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proceedings of Language and Resource Evaluation Conference, LREC 2014*, p. 1267–1274.

GROUIN C. & NÉVÉOL A. (2014). De-Identification of Clinical Notes in French: towards a Protocol for Reference Corpus Developpement. *Journal of Biomedical Informatics*, **50**, 151–61.

HIRSCH J., TANENBAUM J., LIPSKY GORMAN S., LIU C., SCHMITZ E., HASHORVA D., ERVITS A., VAWDREY D., STURM M. & ELHADAD N. (2015). HARVEST, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, **22**(2), 263–74.

LOPER E. & BIRD S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, p. 63–70, Philadelphia, Pennsylvania: Association for Computational Linguistics.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MCCLOSKY D. (2010). *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. PhD thesis, Department of Computer Science, Brown University.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PUSTEJOVSKY J. & STUBBS A. (2011). Increasing Informativeness in Temporal Annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, p. 152–160, Stroudsburg, PA, USA: Association for Computational Linguistics.

SAEED M., VILLARROEL M., REISNER A. T., CLIFFORD G., LEHMAN L.-W., MOODY G., HELDT T., KYAW T. H., MOODY B. & MARK R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, **39**, 952–960.

STYLER IV W., BETHARD S., FINAN S., PALMER M., PRADHAN S., DE GROEN P., ERICKSON B., MILLER T., LIN C., SAVOVA G. & PUSTEJOVSKY J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, **2**, 143–154.

SUN W., RUMSHISKY A. & UZUNER O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, **20**(5), 806–13.