Analyse d'expressions temporelles dans les dossiers électroniques patients

MD. Tapi Nzali[1,2] A. Névéol[1] X. Tannier[1]

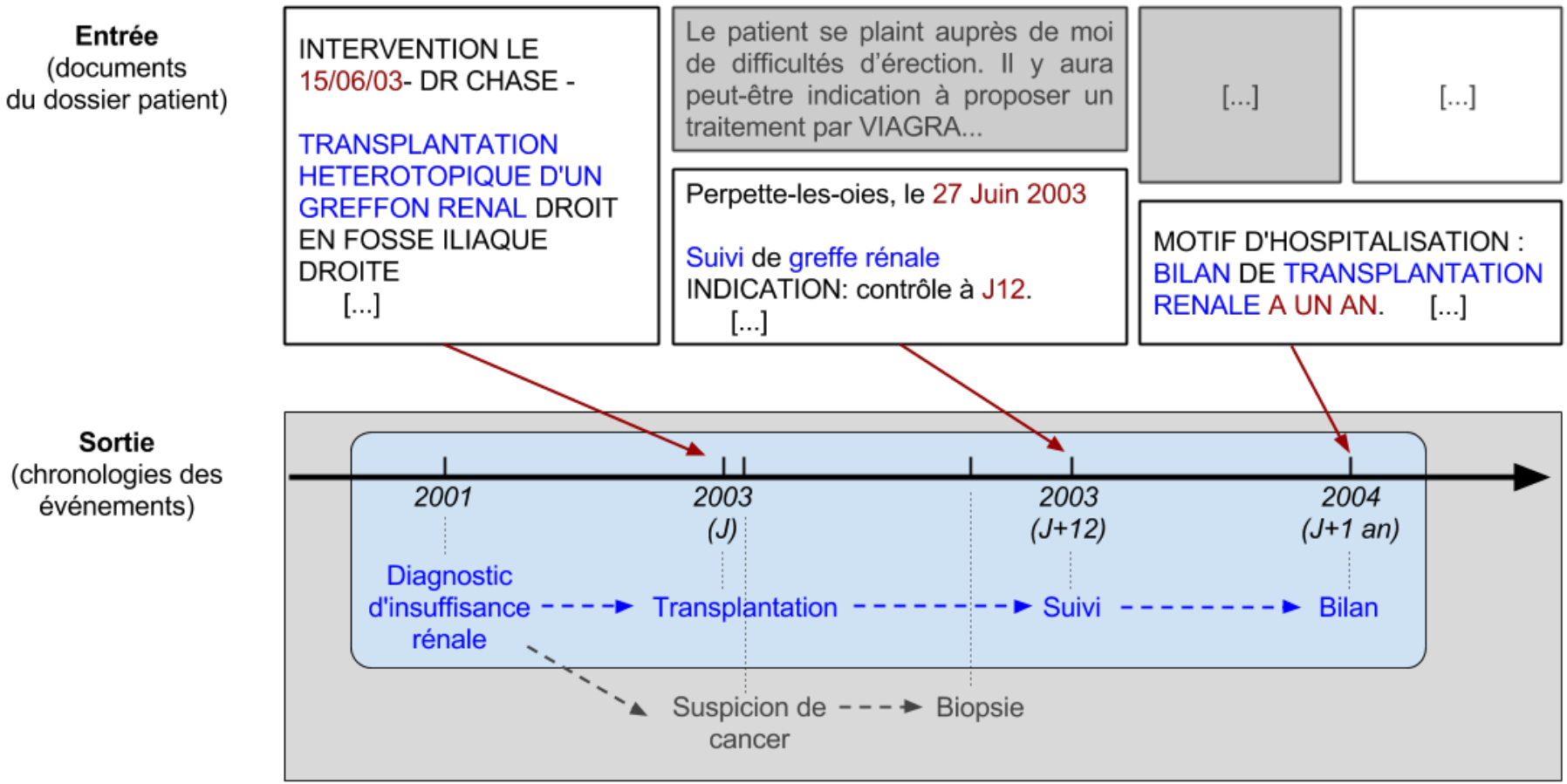
[1]LIMSI
Université Paris-Sud
Orsay, France

[2]LIRMM
Université de Montpellier
Montpellier, France

23 Juin 2015

Plan

Contexte et objectifs



Contexte et objectifs

Objectifs

- Construire la chronologie des événements médicaux d'un patient
- L'analyse rétrospective des cohortes de patients

Première étape

Extraire automatiquement les expressions temporelles et les signaux dans les textes cliniques

Deuxième étape

Détection des événements dans les textes cliniques

Contexte et objectifs

Objectifs

- Construire la chronologie des événements médicaux d'un patient
- L'analyse rétrospective des cohortes de patients

Première étape

Extraire automatiquement les expressions temporelles et les signaux dans les textes cliniques

Deuxième étape

Détection des événements dans les textes cliniques

Mots clés

Dossier patient électronique

Rapports médicaux, hospitalisation, ordonnances, ...

Expressions temporelles

- Date : 03 avril 99, mars 2011, lundi, hier...
- Durée : Suivre le même traitement pendant 10 jours
- Fréquence : prendre 3cp /jour, tous les matin et soir
- Heure : 10h30, midi et demi

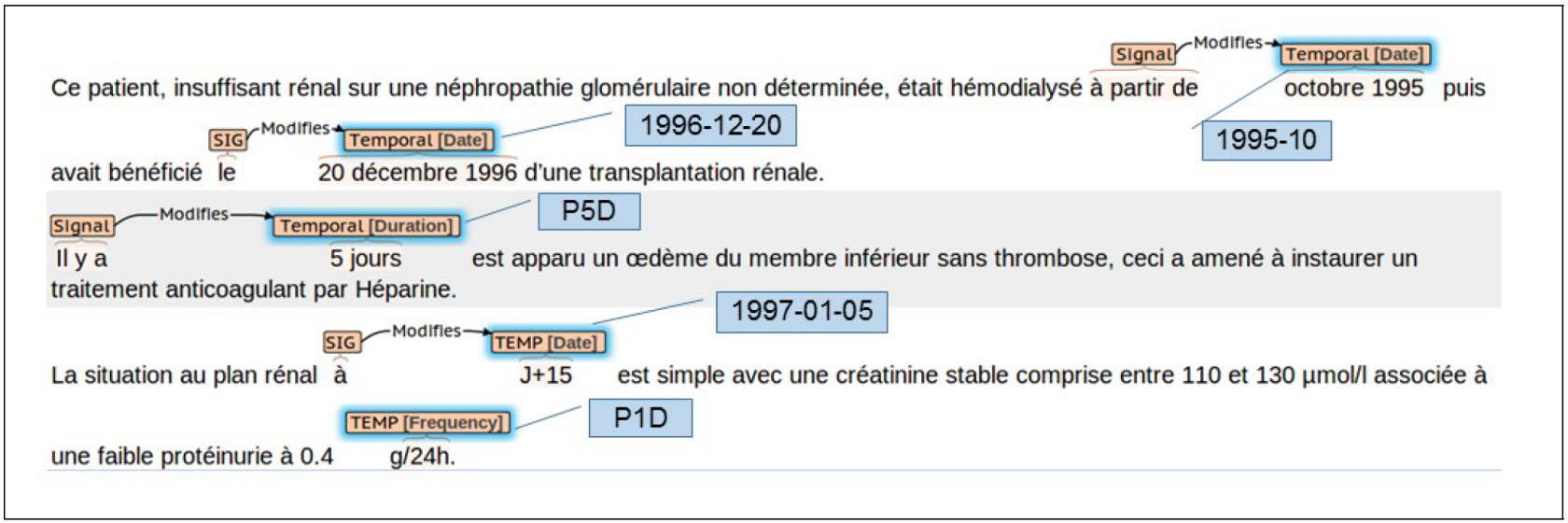
Signaux temporels

- Signal : le, en, pendant ...

Corpus de référence

- Corpus en langue française : 246 documents (training) and 115 documents (test)
- Les renseignements médicaux personnels suppléant (Medina)
- Langage d'annotation : TimeML
- Outil d'annotation : Brat
- 3 annotateurs
- Accord inter-annotateur sur tout le corpus a été de 0,9 de F-mesure

Exemple


Ce patient, insuffisant rénal sur une néphropathie glomérulaire non déterminée, était hémodialysé à partir de
Signal
Modifies
Temporal [Date]
octobre 1995
puis
1995-10
SIG
Modifies
Temporal [Date]
1996-12-20
avait bénéficié le
20 décembre 1996
d'une transplantation rénale.
Signal
Modifies
Temporal [Duration]
P5D
Il y a
5 jours
est apparu un œdème du membre inférieur sans thrombose, ceci a amené à instaurer un traitement anticoagulant par Héparine.
SIG
Modifies
TEMP [Date]
1997-01-05
La situation au plan rénal à
J+15
est simple avec une créatinine stable comprise entre 110 et 130 µmol/l associée à
TEMP [Frequency]
P1D
une faible protéinurie à 0.4
g/24h.

Corpus French TimeBank vs corpus clinique

	FR-TimeBank		Corpus Clinique	
	#	%	#	%
Date	227	53,41%	2594	65,14%
Durée	52	12,24%	343	8,61%
Fréquence	16	3,76%	994	24,96%
Heure	130	30,59%	51	1,28%

HeidelTime modifié (Version française) [X. Tannier, V. Moriceau]

- HeidelTime prend en charge quatre types de base d'objets temporels: **Date, Durée, Heure** and **Fréquence**

Quelques expressions temporelles cliniques

- J5, J-1 pré-op, 5 foisj, 18 semaines d'aménorrhée, à j+1

Règle simple

Extraire et normaliser l'expression temporelle **J+1** dans la phrase *À J+1 post-opératoire, le patient présente de nouveaux symptomes* :
RULENAME="date_r30i",
EXTRACTION="[Jj][\S]*[-][\S]*([\d]+)",
NORM_VALUE="UNDEF-REF-day-MINUS-group(1)"

Un exemple de texte

Ce patient a bénéficié en octobre 1997 d'une transplantation rénale.
A J+1 après l'opération, on a observé une faible protéinurie à 0.4 g/24h.
Traitement:
NEORAL 100 mg le matin, 125 mg le soir
CELLCEPT 2 g par jour

Annotation automatique avec HeidelTime

Ce patient a bénéficié en <TIMEX3 tid="t6" type="DATE" value="1997-10">octobre 1997</TIMEX3> d'une transplantation rénale.
A <TIMEX3 tid="t7" type="DATE" value="XXXX-XX-XX">J+1</TIMEX3> après l'opération, on a observé une faible protéinurie à 0.4 g<TIMEX3 tid="t8" type="SET" value="XXXX-XX-XXT24:00">/24h</TIMEX3>.
Traitement:
NEORAL 100 mg <TIMEX3 tid="t9" type="TIME" value="XXXX-XX-XXTMO">le matin </TIMEX3>, 125 mg <TIMEX3 tid="t10" type="TIME" value="XXXX-XX-XXTEV"> le soir</TIMEX3>
CELLCEPT 2 g <TIMEX3 tid="t11" type="SET" value="xxx" freq="1">par jour</TIMEX3>

HeidelTime modifié

Résultats

	Précision	Rappel	**F-mesure**
Date	0,9144	0,9334	0,9238
Durée	0,8182	0.8090	0,8136
Fréquence	0,4242	0,8209	0,5593
Heure	0,4688	0,0798	0,1364
Global	**0,7666**	**0,7941**	**0,7801**

HeidelTime original	0,5291	0,6283	0,5744
HeidelTime modifié	0,7666	0,7941	0,7801

- Comment augmenter les résultats (F-mesure plus précisement) ?
- Proposition d'un système par apprentissage automatique

Conditional Random Field (CRF) [J. Lafferty, A. McCallum]

Méthode 1

- Prétraitement : Lemmatisation
- Descripteurs : n-grammes de mots, lexiques, patrons syntaxiques, ponctuation, longueur du token, présence d'un chiffre dans le token, capitalisation du token, cluster de brown.
- Classifieur : CRF - Conditional random field
- Validation croisée sur 10 plis
- Outil utilisé : Wapiti [T. Lavergne, O. Cappé, F. Yvon]
- Recherche des meilleurs paramètres de configuration du modèle

Fichier tabulaire : format wapiti

Token	PTT	Length	IsCap	IsPunc	BIO
le	PRP	2	mm	No	B-Signal
17	NUM	2	O	No	B-Date
août	NOM	4	Mm	No	I-Date
1996	NUM	4	O	No	I-Date
,	PUN	1	O	Yes	O

PTT = Part Of Speech TreeTagger

Lenght = Longueur du token

IsCap = IsCapitalized

IsPunct = IsPunctuation

BIO = Begin - Input - Output

Conditional Random Field (CRF)

Résultats

	Précision	**Rappel**	**F-mesure**
Date	0,8246	0,9437	0,8801
Durée	0,6705	0,9833	0,7973
Fréquence	0,8380	0,8693	0,8534
Heure	0,2813	0,9000	0,4286
Global	**0,8068**	**0,9231**	**0,8610**

HeidelTime modifié	0,7666	0,7941	0,7801
CRF	0,8068	0,9231	0,8610

- Peut-on faire mieux ?

Approche hybride

Méthode 2 : HeidelTime modifié + CRF

- Descripteurs de la méthode 1
- Ajout d'un nouveau descripteur : les annotations automatiques faites par "HeidelTime modifié"
- Même principe que la méthode 1

Approche hybride

Résultats

	Précision	**Rappel**	**F-mesure**
Date	0,9262	0,9774	0,9511
Durée	0,7500	0,9296	0,8302
Fréquence	0,8535	0,8601	0,8568
Heure	0,3750	0,8571	0,5217
Global	**0,8837**	**0,9403**	**0,9111**

CRF	0,8068	0,9231	0,8610
Hybride	0,8837	0,9403	0,9111

Évaluation (corpus de test: 115 documents)

	Expressions temporelles (TIMEX)		
	Précision	**Rappel**	**F-Mesure**
HeidelTimeFrench	52,91 %	62,83 %	57,44 %
HeidelTimeFrench (tuning) [HT]	76,66 %	79,41 %	78,01 %
CRF	80,68 %	92,31 %	86,10 %
HT + CRF	**88,37 %**	**94,03 %**	**91,11%**

	Signal		
	Précision	**Rappel**	**F-Mesure**
CRF (signal)	88,50 %	95,66 %	91,94 %
H2 + CRF (signal)	**92,78 %**	**94,04 %**	**93,41 %**

Conclusions

Une tâche qui vise à détecter et classer les dates, les durées, les fréquences, l'heure et les signaux

- Un corpus de référence "clinique" créé par 3 annotateurs
- Modification d'HeidelTime pour identifier les expressions temporelles dans les textes cliniques
- Système d'apprentissage automatique
- Possibilité d'adaptation dans un nouveau domaine et à d'autres tâches.

Perspectives

À court terme

- Normaliser les expressions temporelles

À moyen terme

- Détecter les événements
- Extraire les relations temporelles
- Construire la chronologie des événements du patient

Financement

CABeRneT ANR-13-JS02-0009-01

Merci pour votre attention !