

Le TAL au service du médical

Le projet SYNODOS



André Bittar

Cental, Université catholique de Louvain

29 avril 2016

Contexte

Contexte

Une masse de données médicales

- Les hôpitaux génèrent une quantité énorme de données **textuelles**
- Données largement **non structurées**
- Dossier Patient Informatisé (**DPI**) :
 - Fiche d'identification du patient
 - Motif(s) d'hospitalisation
 - Comptes-rendus opératoires
 - Comptes-rendus d'hospitalisation
 - Conclusions d'examens cliniques successifs
 - Etc.

Contexte

Exploiter automatiquement le DPI

1. Imposer une structure

- Saisie directe dans templates prédéfinis et rigides
- Solution pas optimale :
 - Manque de temps et de motivation
 - Nécessite du texte brut pour des infos particulières aux individus
 - Saisie par diverses personnes (médecins, infirmières, secrétaires, parfois même le patient !)

2. Traiter les données non structurées

- Grand intérêt du TAL
- Deux approches majeures
 - 2.1 Extraction d'informations
 - 2.2 Analyse sémantique

Contexte

Extraction d'informations

- Focalisée sur un aspect / une problématique spécifique
 - Noms de médicaments (Deléger et al. 2010), de gènes (Bunescu et al., 2002), etc.
 - Indicateurs de maladies (Pham et al., 2014)
 - Doses de médicaments (Uzuner et al., 2010)
 - etc.
- Remplissage de templates prédéfinis (ex. i2b2)

Contexte

Analyse sémantique

- Capturer le sens des textes de manière générale
 - Qui ? Quoi ? Quand ? Où ? Pourquoi ? Comment ?
 - Semantic Role Labelling (Gildea & Jurafsky, 2002; Roth et al., 2008; Furstenuau & Lapata, 2009,...)
 - Travaux **récents** sur le français (Pradet, 2015; Danlos et al. 2015)
- Médical
 - Prédominance de l'anglais (Kogan et al., 2005; Yang et al., 2015,...)
 - Relativement peu pour le français (Proux et al., 2012)

Le projet SYNODOS

Le projet SYNODOS

Système de **N**ormalisation et d'**O**rganisation de **D**onnées médicales
textuelles pour l'**O**bservation en **S**anté

- Recherche applicative
- ANR 2012–2015
- Suite d'ALADIN (ANR 2008–2011)
- <http://www.synodos.fr>

Le projet SYNODOS

Objectifs

- Plateforme **générique** d'analyse et structuration **sémantique** des DPI
 - Aide à la décision médicale
 - Surveillance épidémiologique
- **Innovation** : terminologie médicale + analyseur sémantique + système de raisonnement
- Écriture de **règles métier**, indépendantes du domaine de spécialisation
- Deux domaines
 1. Maladies nosocomiales
 2. Délais de prise en charge du cancer du colon

Le projet SYNODOS

Partenaires

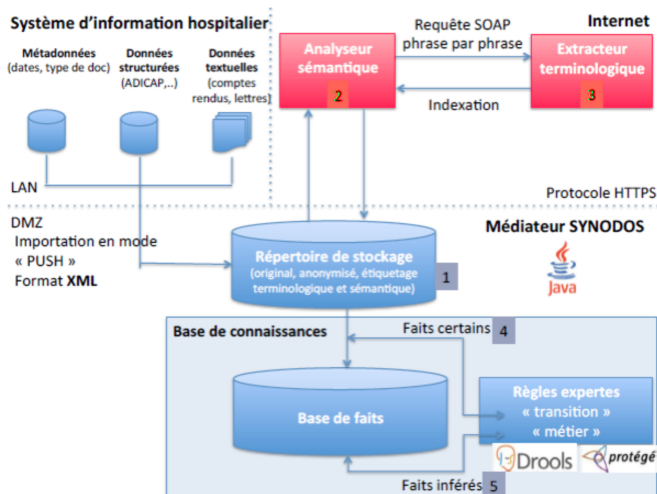
- 2 laboratoires publics
 - LBBE, Laboratoire de biométrie et de biologie évolutive (Lyon) : recherche en épidémiologie
 - CiSMef (Rouen) : recherche en informatique médicale
- 2 industriels
 - Viseo (Grenoble) : “Business Intelligence” et technologies web
 - Holmes Semantic Solutions (Grenoble) : services TAL

Le projet SYNODOS

Holmes Semantic Solutions

- PME : 5 personnes
- Activités commerciales et R&D
- Services TAL
 - Classification de documents
 - Analyse de sentiments & d'émotions
 - Analyse de CV
 - Anonymisation
 - Analyse sémantique générale

Le projet SYNODOS



Le système HOLMES

Le système HOLMES

Hybrid **O**perable platform for **L**anguage **M**anagement and **E**xtensible **S**emantics

- Hybride : symbolique / statistique / apprentissage
- Incrémental : traitements en séquence
- Stanford CoreNLP (Java)

Le système HOLMES

Traitements linguistiques

Segmentation
en phrases

Tokenisation

Etiquetage en
POS

Lemmatisation

Analyse
morphologique

Annotation
lexicale

Grammaires
locales

Analyse
syntaxique

Le système HOLMES

Annotation lexicale

- Gazetteers / lexiques
- Annotations sur séquences de tokens
- Expressions régulières sur **caractères**

Exemple :

```
H([oê]p(ital)?|\.)? (de|\-) la (Croix-Rousse|X\.?R\.)? HOSP
H([oê]p(ital)?|\.)? la (Croix-Rousse|X\.?R\.)? HOSP
H([oê]p(ital)?|\.)? (de|\-) la Croix Rouse HOSP
H([oê]p(ital)?|\.)? la Croix Rouse HOSP
...
```


Le système HOLMES

Grammaires locales

- TokensRegex, Stanford CoreNLP (Chang, 2013)
- Annotations sur séquences de tokens
- Expressions régulières sur **tokens**
- Accès à l'ensemble des annotations sur les tokens (ex. graphie, lemme, POS, NER, etc.)
- cf. Unitex

Exemple :

```
# tous les n? jours / mois / ans
{ pattern: ( [{lemma:tout}] [{pos:DET}] ([{pos:CARD}] | $NUM)?
            $TIMEUNIT ),
  action: ( Annotate($0,ner,"TIMEX3"),
            Annotate($0,normalized,"SET") ),
  result: ( HolmesGroup($0,"SET_2","TIMEX3") ),
  name: "SET_2" }
```

Le système HOLMES

Analyse syntaxique en dépendances

- Français : Talismane (Urieli, 2013)
 - Entraîné sur French TreeBank
 - Dépendances surfaciques (Candito et al., 2009)
 - Pas de contrôle, topicalisation, sujet active/passive etc.
 - Jeu de dépendances restreint : ex. modifieurs 1 dépendance mod
 - Pas optimal pour aller vers la sémantique

Le système HOLMES

Stanford Dependencies

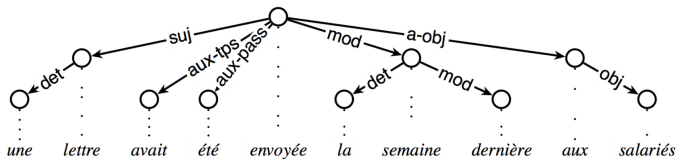
- Représentation syntaxique “profonde” (anglais) (de Marneffe et al., 2008)
- Rapprocher la syntaxe de la sémantique
- Faciliter l'emploi par des non linguistes
- Adaptation au français (Robin, 2013)

Le système HOLMES

Convertir la sortie du parseur

- Dépendances surfaciques $\implies$ SD pour le français

Exemple : Dépendances surfaciques

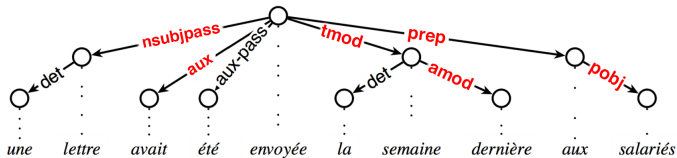


Le système HOLMES

Convertir la sortie du parseur

- Dépendances surfaciques $\implies$ SD pour le français

Exemple : Stanford Dependencies pour le français



Le système HOLMES

Convertir la sortie du parseur

- Transformation de graphe avec SemGreX (Stanford)
- Formalisme déclaratif “maison” pour écriture de grammaires

```
/*  
    Stanford relation pobj  
*/  
68, pobj ([1 pos=P][2 pos§=VINF]) (1-2[type=obj]) =>  
    ([1][2]) (1-2[type=pobj]).  
test:"Jean parle aux salariés."
```

Le système HOLMES

Séquence de grammaires de transformation de graphe

- Conversion dép. surfaciques → SD French
- Identification des pronoms impersonnels (ex. il est ADJ que, il faut/semble/paraît que, il pleut, etc.)
- Traitement de la négation
- **Analyse sémantique**

Données sensibles du DPI

Données sensibles du DPI

Le DPI contient des informations sensibles

- Noms de personnes (patient et sa famille, médecins, etc.)
- Numéro de sécu., téléphone
- Dates (naissance, décès, admission, sortie, etc.)
- Adresses, noms d'établissements

Données sensibles du DPI

Anonymiseur

- Entièrement symbolique / à base de règles
- Gazetteers (prénoms, noms d'établissements, villes de France, etc.)
- Grammaires locales
 - Identification (masquage) des données sensibles
 - Typage :
 - Personne : patient, personnel médical, autre
 - Date : naissance, décès, admission, sortie, autre
 - Lieu : adresse postale, établissement médical, etc.
- Évaluation “informelle” par LBBE

Données sensibles du DPI

Anonymiseur

Exemple :

Madame Dupont, née le 02/07/1956, sera revue en
consultation par le Docteur Martin le 2 février
pour contrôle radio-clinique.

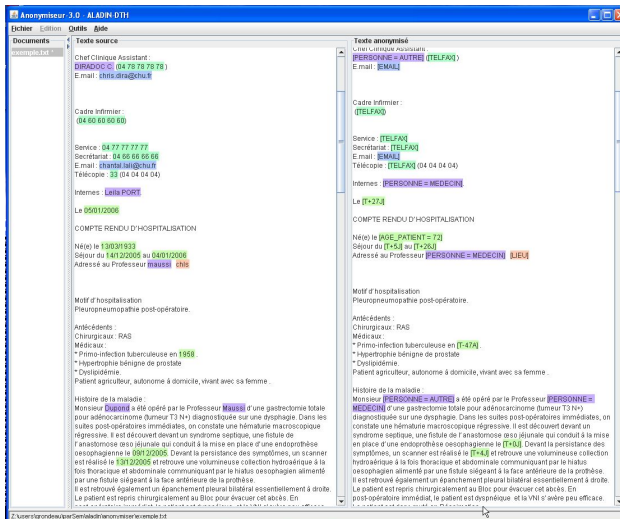
Données sensibles du DPI

Anonymiseur

Exemple :

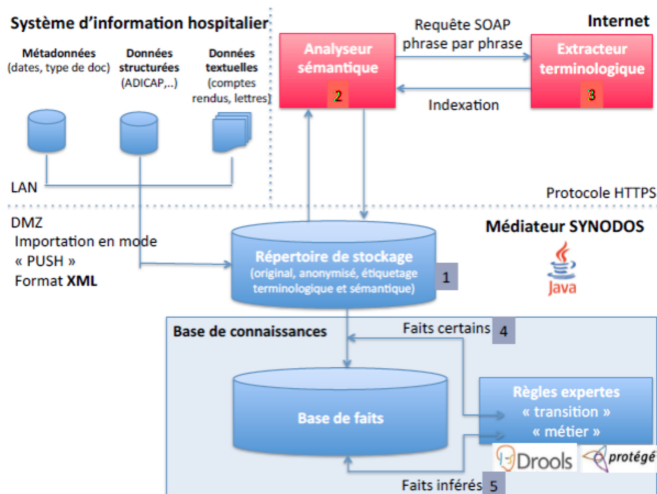
Madame [PERSON=PATIENT], née le [DATE=DOB], sera revue en consultation par le Docteur [PERSON=MED_STAFF] le [DATE] pour contrôle radio-clinique.

Données sensibles du DPI



Analyse sémantique

Analyse sémantique



Analyse sémantique

Annotation terminologique : serveur multi-terminologique du CiSMaF

- 55 terminologies et ontologies médicales : CIM-10, MeSH, ATC, SNOMED, UMLS
- Extracteur de Concepts Multi-Terminologique (ECMT v2)
 - Annotation des concepts médicaux dans une phrase
 - Gestion des variantes morphologiques, syntaxiques, orthographiques

abcès du poumon – abcès pulmonaire

BCPO – bronchopneumopathie chronique obstructive

bronchopneumopathie – broncho-pneumopathie

...

Voir De Meyere et al. (2015)

Analyse sémantique

Annotation terminologique : Serveur multi-terminologique du CiSMeF

- ECMT : ajout de codes des concepts médicaux aux tokens du texte

La voie biliaire du segment II a été cathétérisée.

D340981	D276766	D109888
D384712	D987462	D871666
C0018665	C0087667	C00179774
UMLS_ST_T017	UMLS_ST_T017	UMLS_ST_T058
UMLS_ST_T023	...	...
...		

Analyse sémantique

Annotation terminologique : Serveur multi-terminologique du CiSMeF

- Utilisation des types sémantiques UMLS (Unified Modelling Language System)
- Regroupement des codes en grandes classes

Type d'entité	Types sémantique UMLS
ANIMATED	Organism (T001), Animal (T008), Vertebrate (T010),...
SPATIAL	Spatial Concept (T082), Geographic Area (T083)
MEDICAL ACTIVITY	Health Care Activity (T058), Lab. Procedure (T059),...
PATHOLOGY	Acquired Abnormality (T020), Disease (T047),...
BODY PART	Anatomical Structure (T017), Body Part, Organ (T023),...
CLINICAL DRUG	Antibiotic (T195), Clinical Drug (T200)
RESULT	Finding (T033), Lab. or Test Result (T034)
OBJECT	Physical Object (T072), Manufactured Object (T073),...
FUNCTION	Bio. Function (T038), Phys. Function (T039),...
SUBSTANCE	Chemical (T103), Organic Chemical (T109),...

Analyse sémantique

Annotation terminologique : Serveur multi-terminologique du CiSMeF

Exemple :

La voie biliaire du segment II a été cathétérisée.

[BODY PART]

[BODY PART]

[MEDICAL ACTIVITY]

Analyse sémantique

Semantic role labelling

Quelles sont les relations **sémantiques** entre un prédicat et ses arguments **syntaxiques** ?

Relations sémantiques / rôles thématiques :

- Classiques : AGENT, THEME, EXPERIENCER, LOCALIZATION, MANNER, REASON, etc. (pas PATIENT !) (Qui, où, comment, pourquoi...)
- Temporelles : AFTER, BEFORE, DURING, etc. (Quand)
- Diverses : NEGATION, MODALITY, IMPLICATION, etc.
- Médicales : DOSAGE, CONSUMPTION, MEASUREMENT, PERIODICITY, etc.
- Total : **32**

Analyse sémantique

Semantic role labelling

Quelles sont les relations **sémantiques** entre un prédicat et ses arguments **syntactiques** ?

Cela dépend...

- Propriétés syntactico-sémantiques du prédicat

*Une fiche de liaison est **donnée** à la patiente.* (verbe bitransitif) →
THEME(donnée,fiche) BENEFICIARY(donnée,patiente)

*Le patient est **décédé**.* (verbe inaccusatif) → THEME(décédé,patient)

Analyse sémantique

Semantic role labelling

Quelles sont les relations **sémantiques** entre un prédicat et ses arguments **syntactiques** ?

Cela dépend...

- Type sémantique du prédicat

*Nous **observons** une baisse de TA.* → EXPERIENCER(*observons*,*Nous*)

Radiographie *du genou* (événement) → THEME(*Radiographie*,*genou*)

Abcès *du genou* (objet) → LOCALIZATION(*Abcès*,*genou*)

Analyse sémantique

Semantic role labelling

Quelles sont les relations **sémantiques** entre un prédicat et ses arguments **syntactiques** ?

Cela dépend...

- Types sémantiques des arguments

*Le remplissage a été effectué par **Voluven**.* →
MEANS(remplissage, Voluven)

*Le remplissage a été effectué par **le Docteur Martin**.* →
AGENT(remplissage, Docteur)

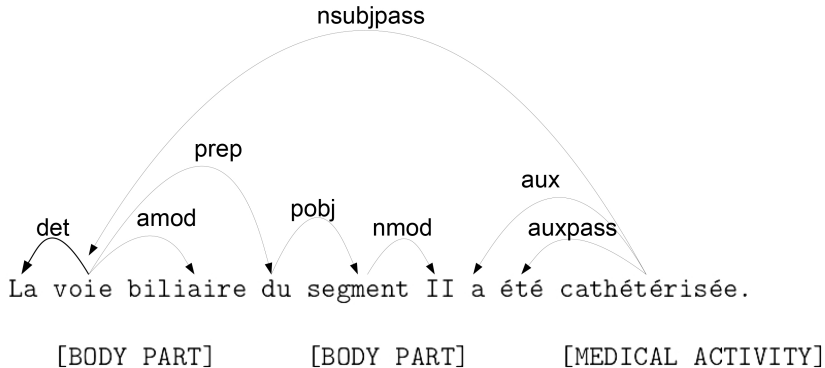
Analyse sémantique

Ressources utilisées

- Annotation terminologique (codes UMLS)
- Ressources lexicales
 - Classifications des verbes, ex. inaccusatifs, bitransitifs – Lefff (Sagot, 2010)
 - Classification de prédicats verbaux et nominaux (ISO-TimeML)
 - Gazetteers entité HUMAN:+ SUMO + WoLF (Sagot & Fišer, 2008)
- Grammaire locale temporelle (ISO-TimeML)
- Grammaires de transformation de graphe : environ 800 règles

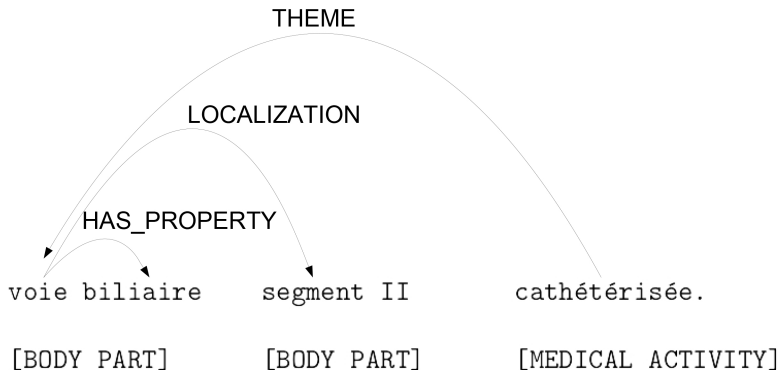
Analyse sémantique

Semantic role labelling



Analyse sémantique

Semantic role labelling



Analyse sémantique

Grammaire syntaxe $\rightarrow$ sémantique

Exemple :

```
/* Agent. */  
23, agent_1 ([1 pos=/V.*/ unacc=FALSE eclass§=/CAUSE|PERCEPTION/  
             [2 anim=TRUE]) (1-2[type=nsubj]) =>  
             ([1][2]) (1-2[type=agent])).  
test:"Jean mange."
```

Évaluation de l'analyseur sémantique

Évaluation de l'analyseur sémantique

Méthodologie

- Guide d'annotation
- Corpus annoté manuellement (gold standard)
- Mesurer :
 - Accord inter-annotateur
 - Performances système
- Analyse d'erreurs

Évaluation de l'analyseur sémantique

Guide d'annotation

- Rédigé en collaboration
- TAListe : aucune connaissance médicale
- Médico-informaticien : connaissances linguistiques limitées
- Consultation, itération et mise à jour

Évaluation de l'analyseur sémantique

Élaboration du gold standard

- Base : 1 900 documents anonymisés
- Sous-ensemble de 12 / 32 relations sémantiques
- Annoter 80 relations de chaque type

Évaluation de l'analyseur sémantique

Annotation

- 2 experts du domaine médical
- 2 annotations indépendantes
- BRAT : **B**rat **R**apid **A**nnotation **T**ool (brat.nlplab.org)
- Échantillonnage aléatoire de textes
- Adjudication, validation → corpus
- Accord inter-annotateur : **0.41** F-score
- Comparaison : **0.47** (Roberts et al., 2008), **0.41** (Deléger et al., 2014), **0.39** (Sun et al., 2012), **0.45** (Styler et al., 2014)

Évaluation de l'analyseur sémantique

Résultats

- BRATEval
- Précision, rappel, F-score
- “Exact match” (entité + relation)

Évaluation de l'analyseur sémantique

Résultats

Relation	P	R	F	Comparaison
NEGATION	0.70	0.63	0.66	0.80 (Mehrabi et al., 2015)
HAS_PROPERTY	0.70	0.50	0.58	
THEME	0.44	0.51	0.47	
AGENT	0.41	0.48	0.44	
ENDS_AT	0.49	0.27	0.34	0.26 (Lin et al. 2016)
BEGINS_AT	0.49	0.23	0.32	0.15 (Lin et al. 2016)
QUANT	0.19	0.35	0.24	0.61 (Lin et al. 2016)
PERIODICITY	0.29	0.20	0.24	
MEASUREMENT	0.21	0.24	0.23	
DURING	0.19	0.21	0.20	
LOCALIZATION	0.13	0.24	0.17	
CAUSE	0.11	0.01	0.02	
Moyenne	0.34	0.34	0.34	

Évaluation de l'analyseur sémantique

Erreurs

- Problèmes d'analyse syntaxique
 - Parseur hors domaine (journalistique versus médical)
 - Mots inconnus
 - Phrases longues et complexes
 - Phrases averbales fréquentes

En somme, patient de [AGE] ans aux antécédents de lymphome malin non Hodgkinien en rémission avec myélodysplasie séquellaire, coronaropathie avancée, insuffisance rénale chronique, admis en Réanimation le [DATE] pour pneumopathie à levures au décours d'une neutropénie associée à une insuffisance cardiaque gauche sur myocardiopathie ischémique.

Évaluation de l'analyseur sémantique

Erreurs

- Manque de connaissances du domaine médical

*Hospitalisation pour décompensation sur une pneumopathie à
Pneumocoque.*

CAUSE(pneumopathie,décompensation)

CAUSE(Pneumocoque,pneumopathie)

Évaluation de l'analyseur sémantique

Erreurs

- Manque de couverture (relations temporelles) :

*Le patient a présenté **dès les premiers jours** des épisodes de désaturation.*

*la prise **pendant une période** de Tazocilline*

- Différences systématiques :

*Le patient est mis au fauteuil depuis **le 13 janvier**. (déterminant inclus / exclus)*

Évaluation de l'analyseur sémantique

Erreurs

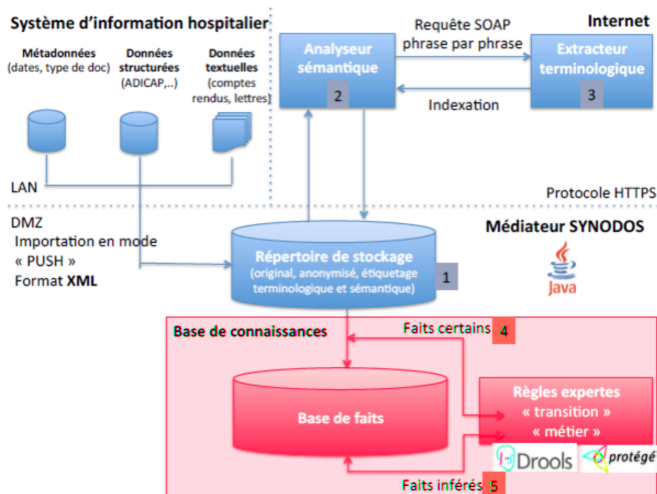
- 18% des erreurs dues à des fautes retrouvées dans le gold standard

*Le patient **ne bénéficie plus** de soin spécifique à la Réanimation.*

NEGATION(soin,plus) / NEGATION(bénéficie,plus)

Systeme expert

Système expert



Système expert

Base de faits – Viseo

- Graphe sémantique pour peuplement de la base de connaissances
- Règles de **transition** : sémantique en faits (faits connus)
- Règles d'**inférence** / **métier** : nouveaux faits (faits inférés)

Système expert

Base de faits – Viseo

Nous avons observé un abcès pulmonaire chez le patient en 2001.

THEME(observé,abcès) DURING(observé,2001)

Règle de **transition** :

if PERCEPTION(e1) & THEME(e1,m1) & DURING(e1,t1) & t1 < REF_TIME **then** (m1, isA, Antecedent)

Règles d'**inférence** / **métier** :

if (m1, isA, Antecedent) & (m1, isA, Disease) **then** (m1, isA, MedicalAntecedent)

if (m1, isA, Antecedent) & (m1, isA, SurgicalProcedure) **then** (m1, isA, SurgicalAntecedent)

Système Expert

- Interface web pour accès aux DPI et écriture de règles métier par les médecins
- Recherche “intelligente” du contenu **sémantique** des textes
- Tests positifs
 1. Maladies nosocomiales
 2. Délais de prise en charge du cancer du colon

Conclusion

Conclusion

SYNODOS

- Réel potentiel de l'analyse sémantique générale pour le médical
- Intérêt d'associer différentes technologies (surtout **TAL** !)
- Première évaluation d'analyseur sémantique médical pour le français
- Besoins TAL
 - Ressources pour le domaine
 - Ressources pour le français
- Explorer approches hybrides : robustesse

Merci !