Vers la simplification de contenus techniques

Natalia Grabar
natalia.grabar@univ-lille3.fr
http://natalia.grabar.free.fr

CNRS UMR8163 STL, Université Lille 3

Octobre 2017

Contexte

- Domaine biomédical :
  - différents types d'utilisateurs
  - experts, patients, pharmaciens, étudiants ...
  - différents niveaux de spécialisation
- Patients : qualité des informations, compréhension
  - Qualité médicale des informations :
    - HONcode éthique : certification des sites de santé
    - autorité, complémentarité, confidentialité, attribution, justification, transparence de financement ...
    - [Risk & Dzenowagis, 2001] : *Review of Internet information quality initiatives*
  - Comfort visuel
  - *Spécialisation technique et scientifique*
  - ...

$\Rightarrow$ Relation directe avec la vie et le bien-être des personnes

- [AMA, 1999, Berland *et al.*, 2001, McCray, 2005, Tran *et al.*, 2009]

FALC : Facile à Lire et à Comprendre

- Objectif :
  - rendre compréhensibles les informations institutionnelles pour
    - le grand public
    - les personnes avec des pathologies
    - les personnes avec le handicap intellectuel
    - ...
- Motivation : législation Européenne en vigueur à partir de janvier 2015
- Quelques exemples :
  - Cochrane [Collaboration, 2009]
    - `http://www.cochranelibrary.com/` (Plain language summary)
  - Encyclopédies en ligne :
    - `https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal`
    - `https://fr.vikidia.org/wiki/Vikidia:Accueil`

Health Literacy / Alphabétisation médicale

- Facilité à lire, comprendre et utiliser les informations de santé
- Dans différents contextes :
  - suivre les instructions de traitement
  - prendre les décisions (maladies chroniques)
  - communiquer avec les médecins
  - réussir le processus de soins
- Tester le niveau de health literacy d'une personne :
  - `https://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/literacy/index.html`

Readability / Lisibilité de documents

- Lisibilité :
  - facilité avec laquelle un document est formulé et écrit
  - propriété d'un texte ou d'un document
    - construction de phrases
    - choix de mots
    - ...
- Tester le niveau de health literacy d'un document :
  - `https://www.webpagefx.com/tools/read-able/`
  - `https://readable.io/`

Lisibilité de documents médicaux

- Une réelle difficulté de compréhension :
  - compréhension des différentes étapes pour la bonne administration de médicaments [Patel *et al.*, 2002]
  - cohorte de 2 600 patients américains (2 hôpitaux) :
    - entre 26 % et 60 % ne peuvent pas comprendre les instructions sur les médicaments, les consensus informés, les brochures de santé [Williams *et al.*, 1995]
  - documents, sites web de santé à destination des patients :
    - montrent souvent des niveaux de spécialisation élevés [Berland *et al.*, 2001]

Ce que cela donne du côté des patients...

[Guilbert, 2014]

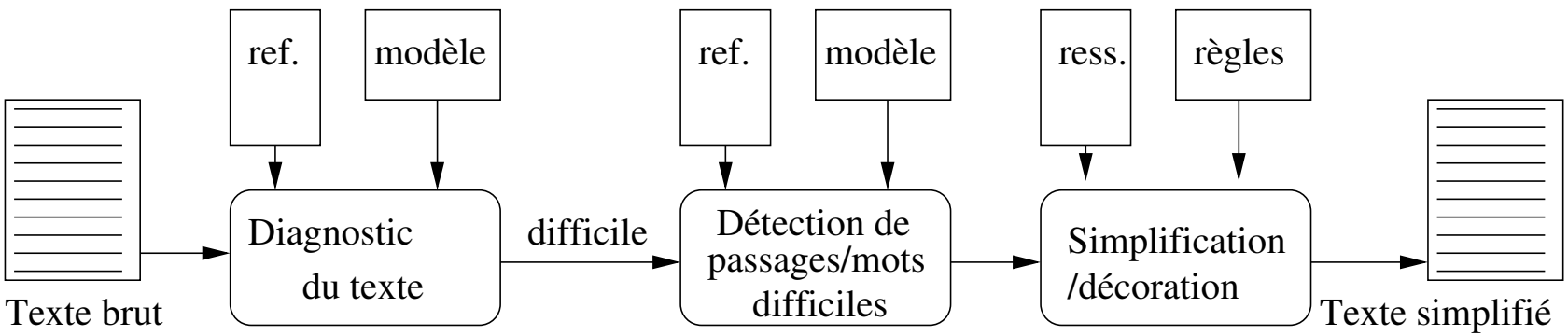
- *Docteur, j'ai une hernie fiscale*
  → *...hernie discale*
- *Docteur, j'ai une fuite mistrale*
  → *...fuite mitrale*
- *J'ai dû subir une enculoscopie*
  → *...coloscopie*
- *J'ai fait un coma idyllique*
  → *...coma éthylique*
- *J'ai consulté un gastro-entéropode*
  → *...gastro-entérologue*
- *On m'a fait 3 points de soudure*
  → *...suture*
- *J'ai entendu à la radio que vous pouviez me donner des gélules souches*
- *J'ai une augmentation des triglicirilіques*
- *Faut m'opérer du corps vitreux*

ETP : éducation thérapeutique des patients

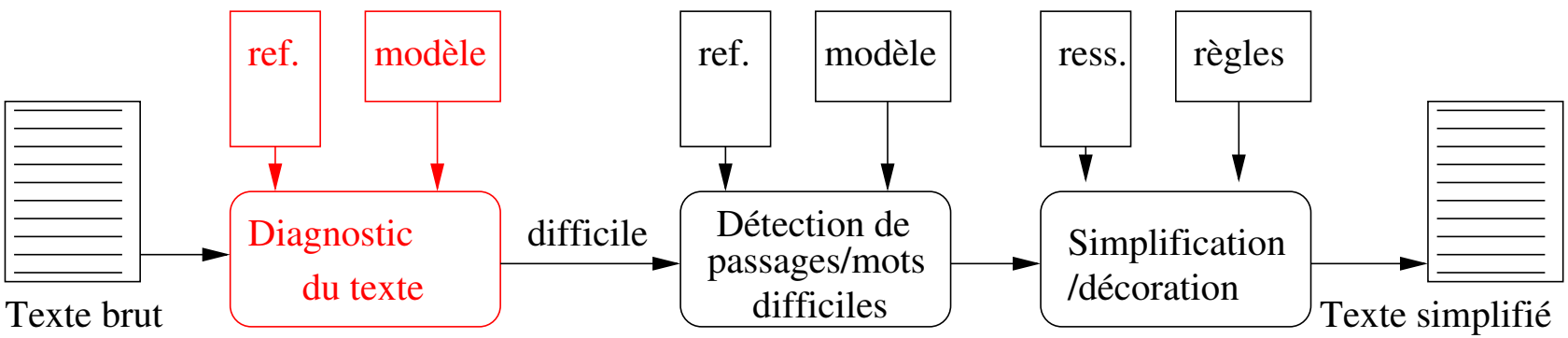
- Objectif [Golay *et al.*, 2007, Glasgow *et al.*, 2012] :
  - répondre aux priorités politiques de la santé publique et domaine médical
- Aider les patients avec des pathologies chroniques :
  - acquérir et maintenir le savoir-faire
  - mieux gérer la maladie au quotidien
- Aider les professionels médicaux [d'Ivernois *et al.*, 2011, Gross & Gagnayre, 2013, Brin-Henry, 2014] :
  - mieux communiquer avec les patients
  - mieux guider les patients dans leurs parcours médical
- Établir une confiance mutuelle [Sørensen, 1996]
- Améliorer l'efficacité des soins médicaux

Objectifs

- Rendre les documents de santé mieux compréhensibles par les patients



Diagnostic de la difficulté du texte



Diagnostic de la difficulté du texte

Travaux existants

- Formules de lisibilité : longueur moyenne des mots et phrases
  [Flesch, 1948, Gunning, 1973,
  Björnsson & Härd af Segerstad, 1979]
- Formules de lisibilité et vocabulaire médical
  - [Kokkinakis & Toporowska Gronostaj, 2006]
- Apprentissage supervisé avec différents descripteurs
  - [Poprat *et al.*, 2006, Zheng *et al.*, 2002, Grabar *et al.*, 2007,
    Goeuriot *et al.*, 2007, Miller *et al.*, 2007,
    Chmielik & Grabar, 2011]
- Combinaison de différents descripteurs
  - linguistiques, lisibilité, lexique [Wang, 2006]
  - lisibilité, catégories grammaticales, familiarité des termes
    [Zeng-Treiler *et al.*, 2007]
- Étude de la syntaxe [Zeng-Treiler *et al.*, 2007]

Diagnostic de la difficulté du texte

Hypothèses

- Contenu des documents discriminant de leur typologie
  - auteurs, destinataires,...
- Descripteurs linguistiques de différents niveaux :
  - lexical
  - stylistique
  - morphologie
  - ...
- Transposition de descripteurs d'une langue dans une autre

Diagnostic de la difficulté du texte

Expériences

1. Descripteurs morphologiques
2. Descripteurs de la subjectivité
3. Descripteurs discursifs translangues
4. Descripteurs cross-corpus

Descripteurs morphologiques

[Chmielik & Grabar, 2011]

- Langue : français
- Source des documents : portail CISMeF
- Thématiques : cardiologie, pneumologie, hématologie
- Discours : expert, didactique, non expert

	Cardio		Pneumo		Hémato	
	doc.	occ.	doc.	occ.	doc.	occ.
Expert	2 922	1 285 665	1 823	1 265 726	1 580	1 512 064
Didactique	582	384 550	304	229 639	293	198 264
Vulgarisé	404	253 402	317	189 205	203	100 126
Total	3 908	1 923 617	2 444	1 684 570	2 076	1 810 454

- Objectif : catégorisation des documents selon les trois discours
  - exploitation de descripteurs du niveau morphologique

Descripteurs morphologiques

- `TreeTagger` : étiquetage morpho-syntaxique [Schmid, 1994]

est	VER :pres	être
antiinflammatoire	PRO :POS	antiinflammatoire

- `FLEMM` : lemmatiseur et correcteur morphologique [Namer, 2000]

est	VER(pres) :3p :s :pst :ind	être :3g
antiinflammatoire	NOM :_ :s	antiinflammatoire

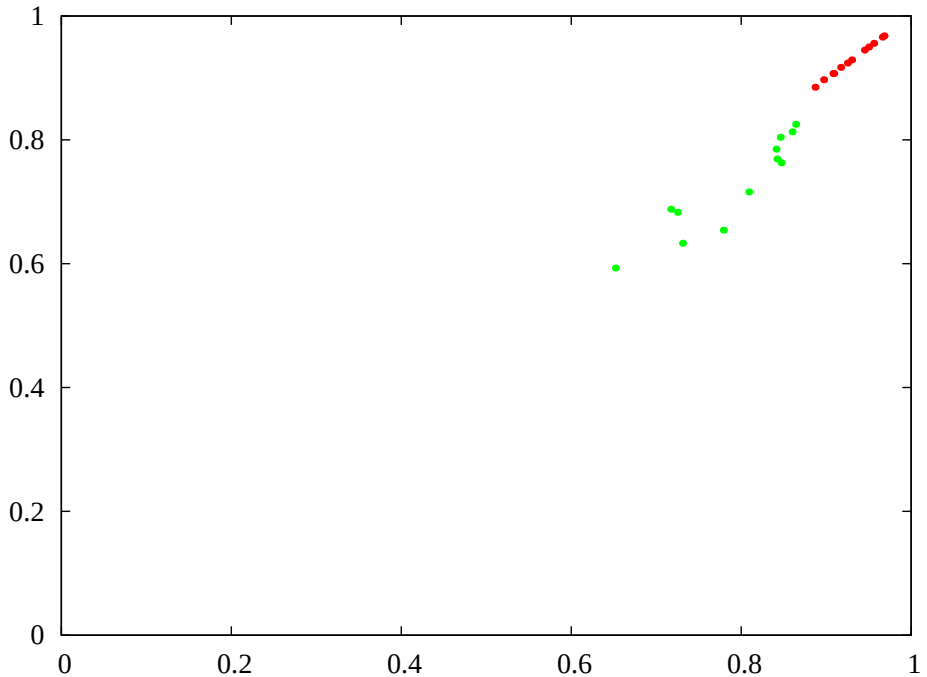
- `DériF` : analyseur morpho-sémantique [Namer, 2003]

  angioblastique/ADJ
  [[angi N*] [blast N*] ique ADJ]
  (angioblastique/ADJ, [angi,N*] :blast/N*)
  Qui est en relation avec cellule embryonnaire et vaisseau
  Constituants = /angi/blast/ique
  Type = anatomie

Descripteurs morphologiques

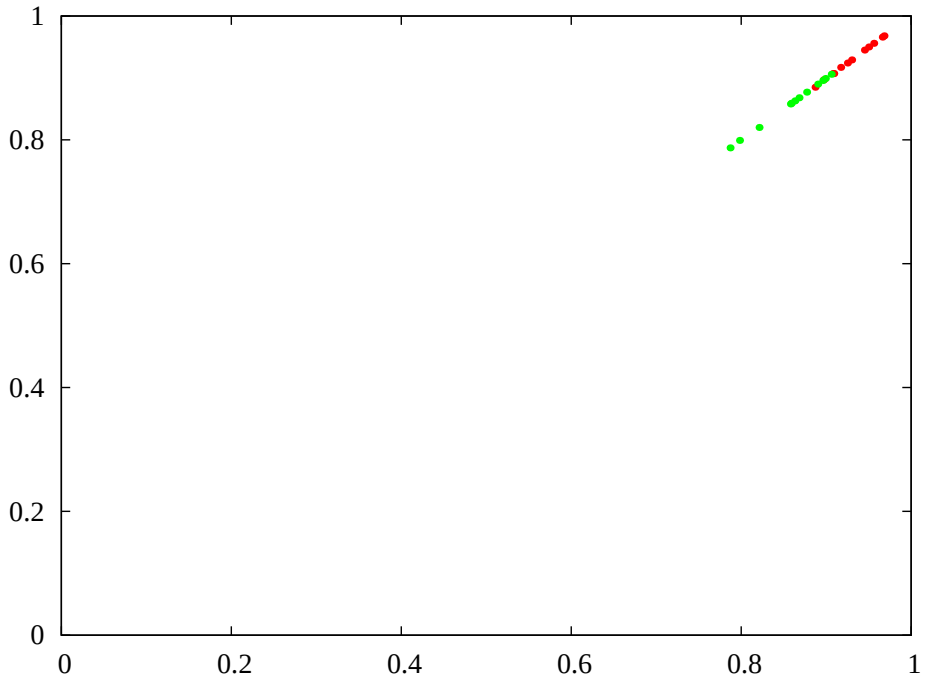
- Apprentissage supervisé : *Support Vector Machine*
- Trois catégories : pro, etu, vul
  - bi-catégorie et tri-catégorie
- Descripteurs : bases *b*, bases et affixes *ba*
  - *b* : *angi*, *blast*
  - *ba* : *angi*, *blast*, *ique*
- Présence et fréquence des descripteurs
- Pondération : *freq*, *norm*, *tfidf*
- Sélection : nombre minimum de documents [1 ; 10]
- Baseline : lexique lemmatisé

Descripteurs morphologiques



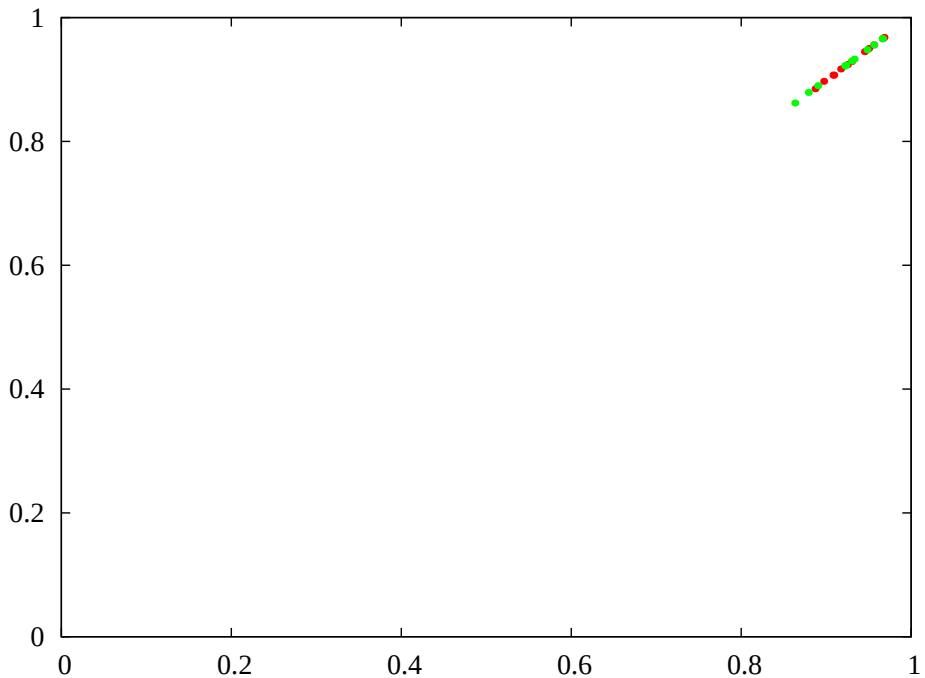
- Unités morphologiques plus discriminantes que la baseline

Descripteurs morphologiques



- Présence et fréquence
  de lexèmes répondant aux procédés morphologiques

Descripteurs morphologiques



- Bases et affixes et bases seulement

Descripteurs morphologiques

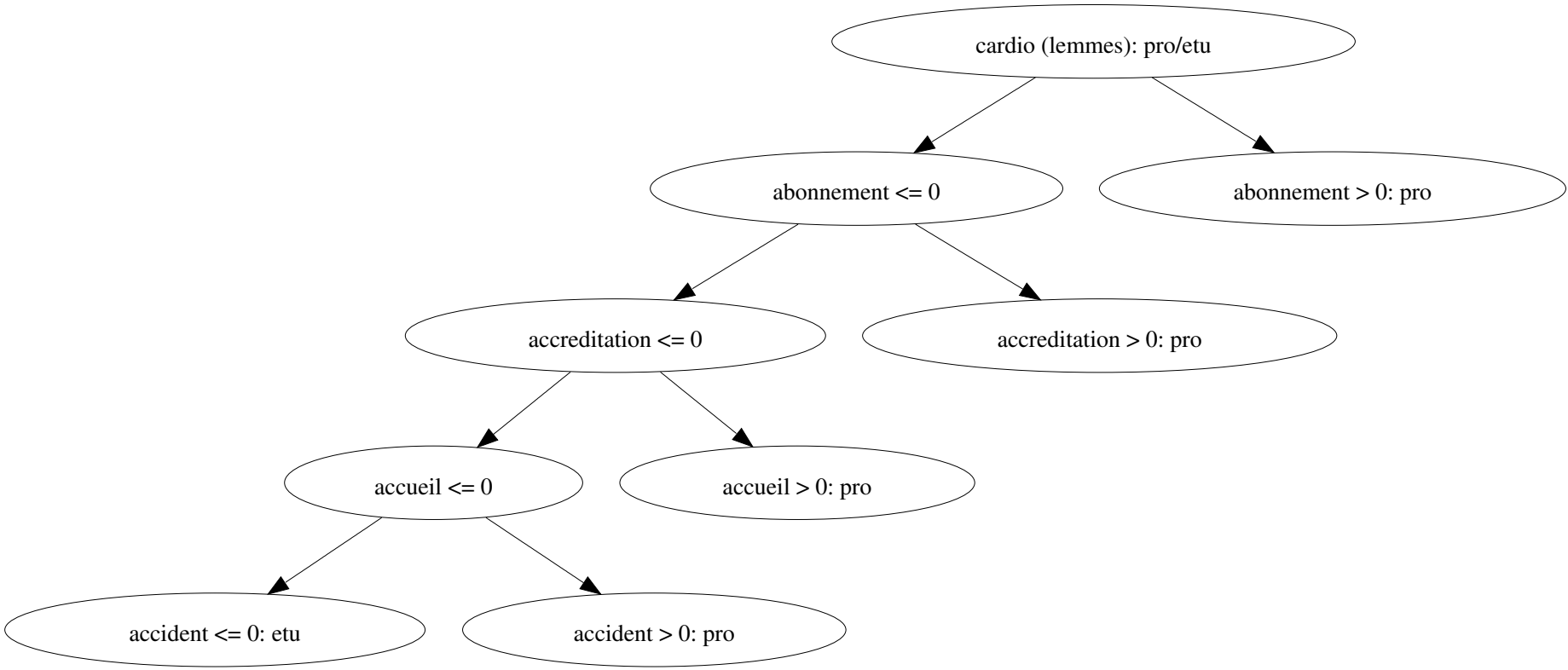
Les descripteurs les plus fréquents (bases et affixes) :

- *pro* : *ion al is logie able techno ité alerte économie acte organe nation mentir*
- *étu* : *ion patho traiter ose graphie génèr cardia thérapie isch thromb pulmon*
- *vul* : *ion ique ité cardia prévent utile traiter infect guider informe allergie post*

Hapax

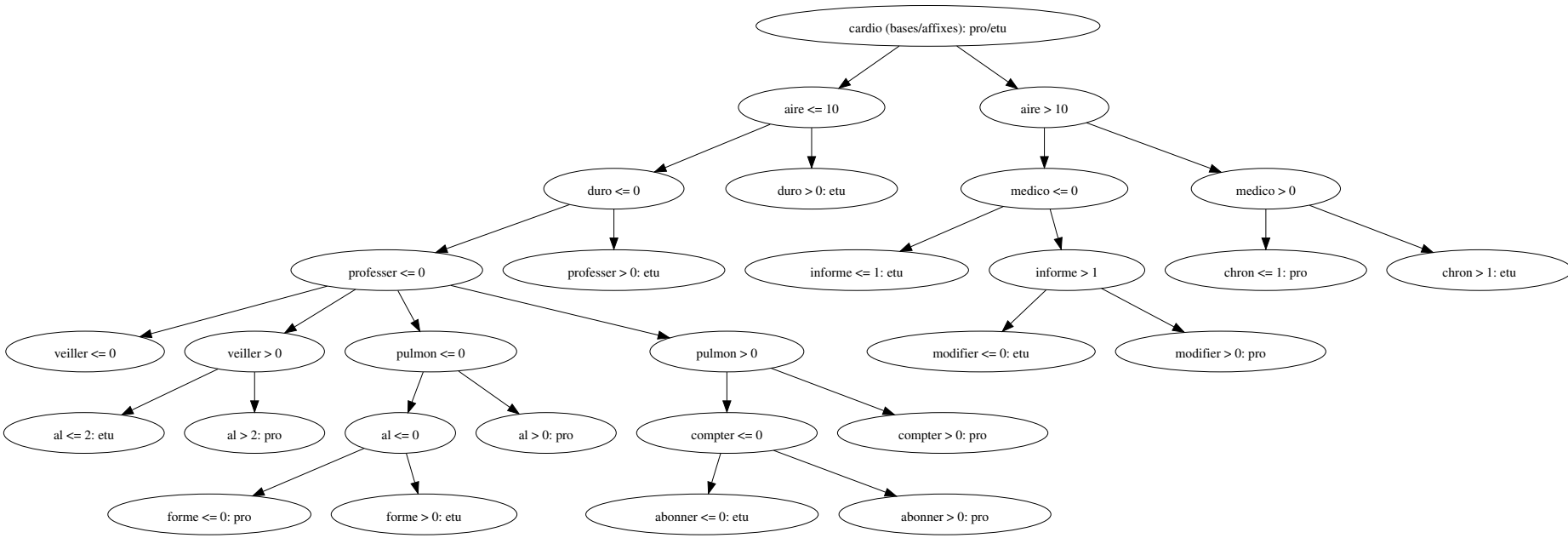
corpus	total	uniq	exemples
cardio	4 372	1 889	$jéjuno_e$ $plausible_e$ $flavone_p$ $ischio_p$ $méiose_p$ $normat_v$
pneumo	4 260	1 899	$abort_e$ $alopécie_e$ $amiodarone_p$ $monét_p$ $spéc_v$ $fécal_v$
hémato	4 097	1 877	$abdomen_e$ $naevo_e$ $cocaïne_p$ $phrén_p$ $angél_v$ $règle_v$
pro	4 887	1 786	$adipos_c$ $brachy_c$ $ankylo_h$ $carotène_h$ $agrégat_p$ $aphte_p$
étu	3 926	1 285	$abduct_c$ $crano_c$ $mnés_c$ $coccygien_h$ $exérèse_p$ $ptérygp$
vul	2 645	1 119	$adip_c$ $graphe_c$ $pexie_c$ $amnio_h$ $angél_h$ $cili_p$ $gnath_p$

Descripteurs morphologiques



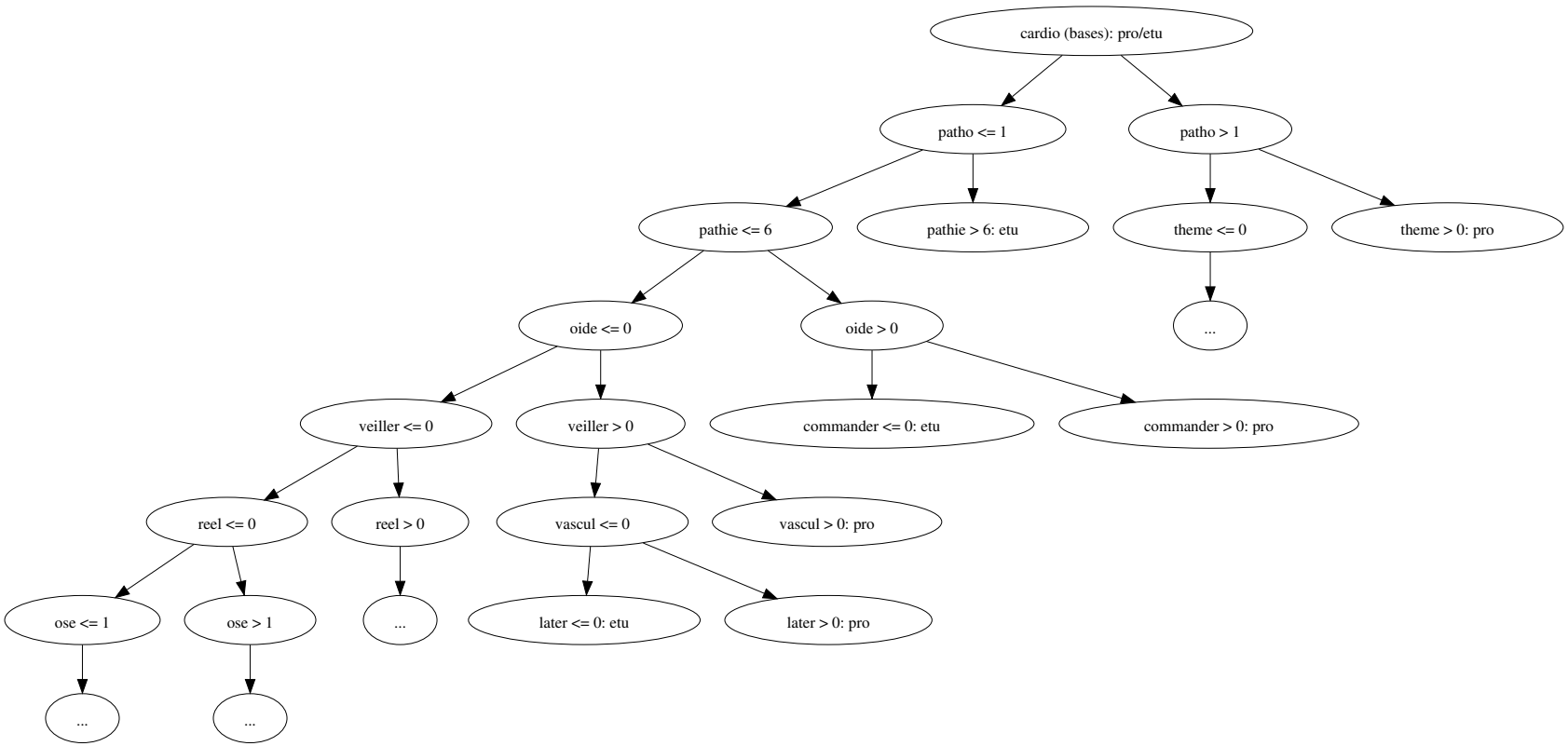
- cardiologie, fréquences, *pe*, lexique, 10 occ

Descripteurs morphologiques



- cardiologie, fréquences, *pe*, bases/affixes, 10 occ

Descripteurs morphologiques



- cardiologie, fréquences, *pe*, bases, 10 occ

Diagnostic de la difficulté du texte

Expériences

1. Descripteurs morphologiques
2. Descripteurs de la subjectivité
3. Descripteurs discursifs translangues
4. Descripteurs cross-corpus

Descripteurs de la subjectivité

[Grabar *et al.*, 2015]

- Langue : français
- Source des documents : CISMeF, clinique, forum
- Thématique : rhumatologie
- Discours : scientifique, clinique, patient

	Rhumatologie		
	doc.	occ.	moyenne
Scientifique	265	840 228	3 170
Clinique	8 162	5 806 158	711
Forum	4 388	3 351 951	763
Total	12 815	9 998 337	780

- Objectif : exploiter la subjectivité pour la catégorisation

Descripteurs de la subjectivité

- `TreeTagger` : étiquetage morpho-syntaxique [Schmid, 1994]

est	VER :pres	être
antiinflammatoire	PRO :POS	antiinflammatoire

- Annotation sémantique avec Snomed Int
  - maladies, actes médicaux, traitements
  - erreurs d'orthographe [Levenshtein, 1966]
- Détection de la négation (*absence, pas de, non*) et de l'incertitude (*possible, incertain, peut être*)
- Détection des modifieurs (*très, peu*)
- Détection des émotions :
  - lexical [Augustyn *et al.*, 2008] :
    gorge irritée, jambe tendue, manque de preuve
  - smileys : *=), ;-), :-/, XD*,
  - typographique : *lol*, *mdr*, *haha*, *hihi*
  - ponctuations expressives : *!!!??*, *!!!!!!!!!!*;
  - mots avec lettres répétées : *maaaaaaal*

Descripteurs de la subjectivité

- Apprentissage supervisé : *RandomForest*
- Trois catégories : scientifique, clinique, forum
  - bi-catégorie et tri-catégorie
- Pondération : *freq*, *norm*, *tfidf*
- Baseline : assignation dans la catégorie par défaut

Descripteurs de la subjectivité

Évaluation de l'annotation sémantique :

- précision stricte P_s (détection, typage et lemmatisation corrects)
- précision lâche P_l (détection correcte)

Corpus	P_s	P_l
Clinique	0,87	0,91
Scientifique	0,81	0,89
Forum	0,88	0,90
Moyenne	0,85	0,90

Descripteurs de la subjectivité

Évaluation de la catégorisation automatique

Catégories	freq	norm	tfidf
Clin./Forum	0,937	0,948	0,946
Forum/Scient.	0,936	0,928	0,940
Clin./Scient.	0,909	0,946	0,911
Clin./Scient./Forum	0,891	0,903	0,877

- Gain par rapport à la *baseline* :
  - entre 0,818 et 0,896 pour les tests à deux catégories
  - entre 0,824 et 0,861 pour le test à trois catégories
- Descripteurs saillants :
  - scientifique : maladies, traitements, incertitude
  - clinique : actes médicaux, traitements, négation
  - forum : émotions, maladies, erreurs d'orthographe

Diagnostic de la difficulté du texte

Expériences

1. Descripteurs morphologiques
2. Descripteurs de la subjectivité
3. Descripteurs discursifs translangues
4. Descripteurs cross-corpus

Descripteurs discursifs translangues

[Grabar & Krivine, 2007, Grabar *et al.*, 2007]

- Corpus source : russe, *diabète et diète*
- Corpus cible : français, *pneumologie*
- Source des documents : web, CISMeF
- Discours : expert, non-expert

	Diabète		
	sites	doc.	occ.
Expert_{ru}	21	35	116 000
Non-expert_{ru}	52	133	190 000
Expert_{fr}	46	186	371 045
Non-expert_{fr}	31	80	87 177

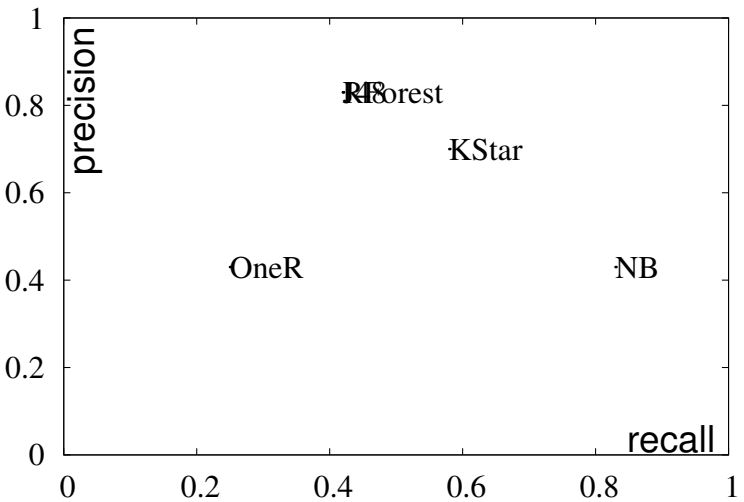
- Objectifs : expérience translangue de catégorisation des documents
  - descripteurs discursifs

Descripteurs discursifs translangues

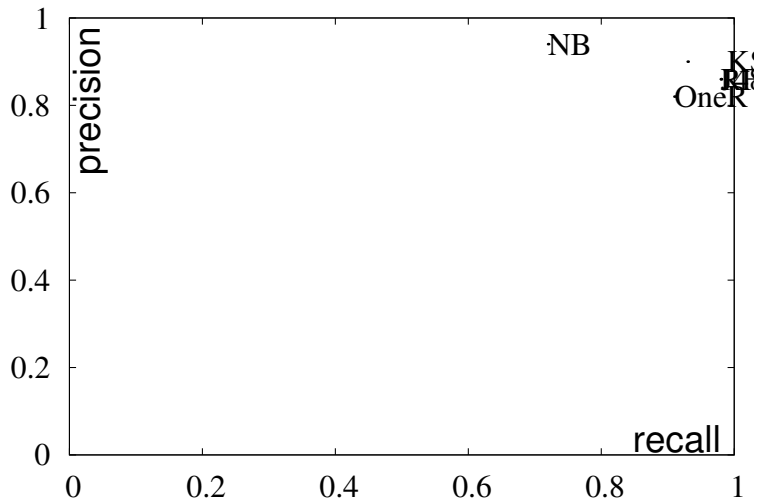
- Apprentissage supervisé avec Weka [Witten & Frank, 2005]
- Algorithmes de différentes familles
- Descripteurs :
  - Structure et mise en page des documents
    - images, tables, listes, éléments des listes, hyperliens, italics, gras
  - Pronoms personnels :
    - 1^{ere} et 2^{ere} personnes au singulier et au pluriel
  - Ponctuation : point d'exclamation (*!*)
  - Incertitude : бы, conditionnel (*rrait*, *raient*...)
- Évaluation
  - Apprentissage et test sur des corpus indépendants
  - 66 % et 33 % des corpus
  - Précision et rappel

Descripteurs discursifs translangues

Corpus russe (source)



Corpus expert

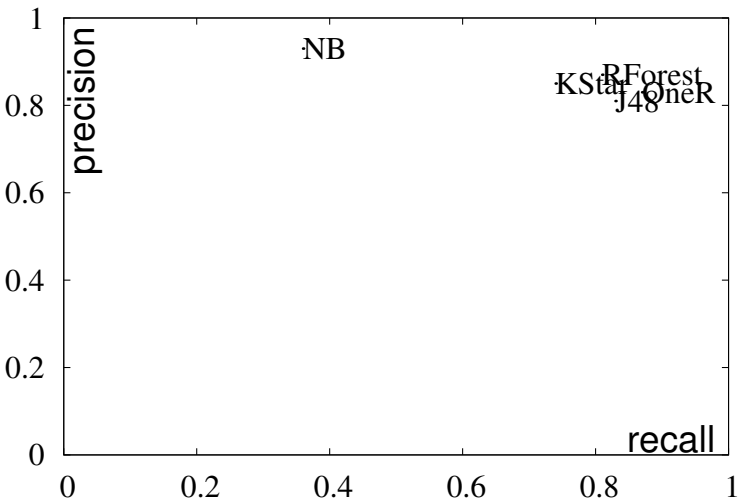


Corpus non expert

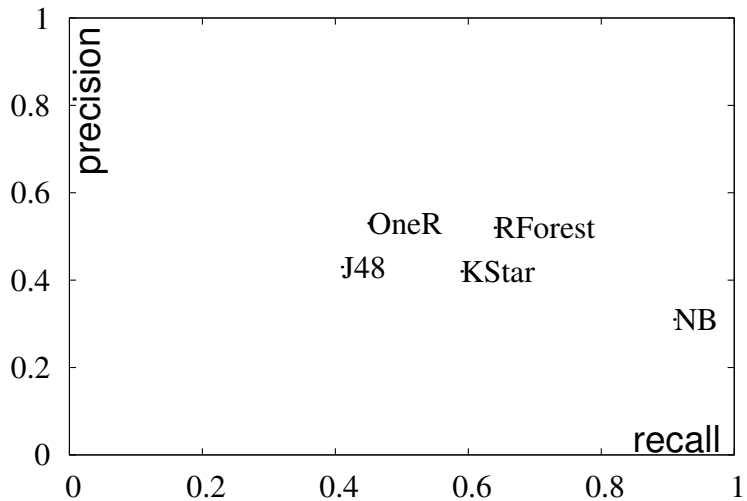
- Documents non expert : mieux reconnus
- Meilleurs algorithmes :
  arbres de décision (J48, RandomForest), *lazy* (KStar)

Descripteurs discursifs translangues

Corpus français (cible)



Expert corpus

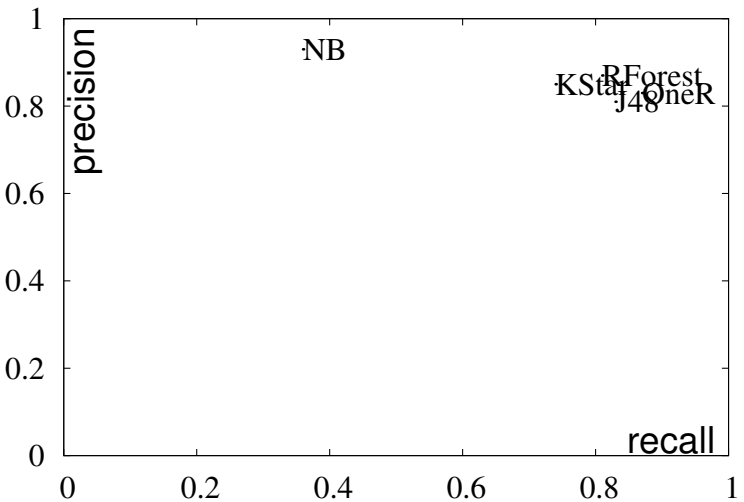


Non expert corpus

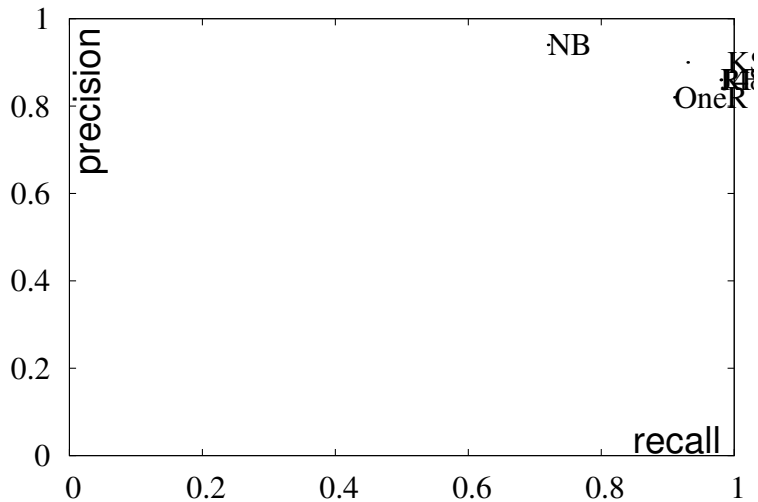
- Documents expert : mieux reconnus
- Meilleurs algorithmes :
  arbres de décision (RandomForest)

Descripteurs discursifs translangues

Sensibilité à la taille de corpus



Français expert



Russe non expert

- Meilleures performances avec un grand corpus

Descripteurs discursifs translangues

medi.ru »» Фармацевтические компании »» Ново Нордиск »» Сахарный диабет I типа »» Новости диабетологии

Информация для профессионалов здравоохранения ! Соглашение об использован

Управление диабетом детей и подростков
(руководство для обучающих управлению диабетом)

ОБУЧАЮЩЕЕ РУКОВОДСТВО ДЛЯ ДЕТЕЙ И ИХ РОДИТЕЛЕЙ ДЕТСКОЕ ОТДЕЛЕНИЕ, КЛИНИКА УНИВЕРСИТЕТА ГЛОСТРУП ДАНИЯ, 1999

Бирте С. Ольсен, консультант-педиатр;
Хенрик Мортенсен, главный врач, старший детский эндокринолог;
Медицинские сестры по диабету Лене Повлсен и Кристен Дюрлов

[Содержание]

ДИАБЕТ И АЛКОГОЛЬ

ИНФОРМАЦИЯ ДЛЯ СПЕЦИАЛИСТОВ

Люди с диабетом должны принимать меры предосторожности во время приема алкоголя. Без специальных знаний и аккуратного планирования могут возникнуть опасные ситуации. Симптомы низкого сахара крови часто ошибочно принимаются за опьянение и поэтому не распознаются и не лечатся. По этой причине важно не пить слишком много. Друзья должны знать об особой опасности алкоголя для людей с диабетом. Люди с диабетом должны всегда носить диабетическую идентификационную карточку.

Сахар крови

Важно проверять сахар крови перед сном после приема алкоголя. Прием алкоголя повышает риск гипогликемии в последующие 24 часа. Этот риск повышается из-за тенденции заменять алкоголем еду и напитки.

ИНФОРМАЦИЯ ДЛЯ СПЕЦИАЛИСТОВ

физическая активность может также повышаться во время приема алкоголя. Например, прием алкоголя часто сопровождается танцами или бодрствованием позже обычного. Кроме того, печень занята тем, что разрушает алкоголь, и ее функция повышать сахар крови не выполняется. Поэтому инъекции глюкагона будут неэффективны, когда в крови имеется алкоголь.

Еда и напитки

- *Non expert* (NaiveBayes, RForest, OneR)
- *Expert* (J48, KStar)

Diagnostic de la difficulté du texte

Expériences

1. Descripteurs morphologiques
2. Descripteurs de la subjectivité
3. Descripteurs discursifs translangues
4. Descripteurs cross-corpus

Descripteurs cross-corpus

[Abdaoui *et al.*, 2014]

- Langue : français
- Source des documents : forums modérés
- Thématiques : divers
- Discours : non expert (questions), expert (réponses)

	messages
AlloDocteurs	4 400
MaSanteNet	12 000
Total	16 000

- Objectif : catégorisation des documents selon les deux discours
  - exploitation du modèle appris sur un autre corpus

Descripteurs cross-corpus

- Plusieurs algorithmes de Weka : SVM, RandomForest, J48, JRip
- Unité : message
- Descripteurs :
  - BOW
  - dictionnaires (termes, émotions, certitude, négation...)
  - BOW + dictionnaires

Descripteurs cross-corpus

Descripteurs	SMO	J48	RandomForest	JRip
BOW	92	90,6	92,1	89,7
Dictionnaires	71,6	73	74	75
BOW+Dictionnaires	92,7	90,7	92,7	90,3

- 10-fold cross-validation sur AlloDocteur

Descripteurs cross-corpus

Descripteurs	SMO	J48	RandomForest	JRip
BOW	100	100	100	100
Dictionnaires	88,9	91,6	93,6	92
BOW+Dictionnaires	100	100	100	100

- 10-fold cross-validation sur MaSanteNet

Descripteurs cross-corpus

Descripteurs	SMO	J48	RandomForest	JRip
BOW	96,6	97,7	98	96,9
Dictionnaires	57	62,1	69,6	69,6
BOW+Dictionnaires	96	97,3	98,2	96,6

- Modèle appris sur AlloDocteurs appliqué sur MaSanteNet

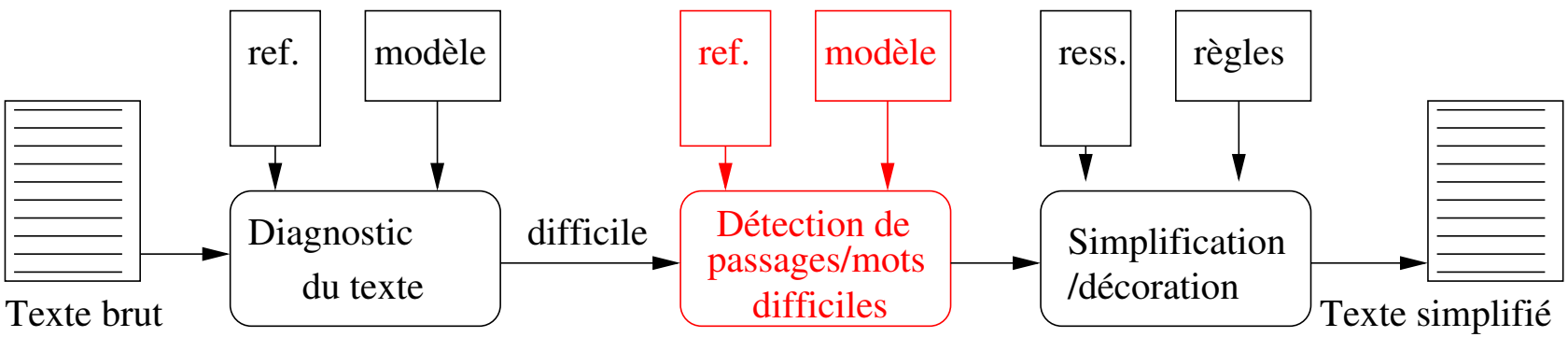
Descripteurs cross-corpus

Descripteurs	SMO	J48	RandomForest	JRip
BOW	37,3	33,3	46,3	33,3
Dictionnaires	57,1	52,9	53,2	55,3
BOW+Dictionnaires	37,5	33,3	43,7	33,3

- Modèle appris sur MaSanteNet appliqué sur AlloDocteurs

Détection de mots/passages difficiles

1. Contexte
2. Diagnostic de la difficulté du texte
3. Détection de mots/passages difficiles
   (Grabar et al., 2014)
4. Acquisition de ressources pour la simplification
5. Simplication/décoration de textes
6. Conclusion



Détection de mots/passages difficiles

Objectifs : détecter les mots difficiles à comprendre

Histoire de la maladie
Le patient a été hospitalisé le 18/7/11 à [redacted] pour un AVC ischémique dans le territoire profond de l'artère cérébrale postérieure droite, thrombolysé à H+3.

Le patient présente, comme déficit, une hypoesthésie gauche et une parésie gauche (force motrice à 1/5 au membre supérieur gauche et 2/5 au membre inférieur gauche), un NIHSS à 8, une désorientation tempora-spatiale et une vigilance fluctuante.
Dans les suites, est survenu un OAP post thrombolyse, probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse).

Le patient est donc transféré en réanimation : l'OAP est résolutif sous VNI et oxygénothérapie.

La majoration de l'insuffisance rénale nécessite 2 cures de dialyse. Mr K. est ensuite transféré en post-réanimation devant l'évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge.

Le 11/8/2011, le patient présente une douleur thoracique associée à une désaturation à 83 %, il est donc transféré en Unité de soins intensifs cardiologiques. Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire. Une anticoagulation curative par CALCIPARINE est mise en place.

Détection de mots/passages difficiles

Matériel

- Objet : termes médicaux (151 104)
- Source : Snomed International [Côté *et al.*, 1993]
- Unité : mots (29 641)
  - lemmes (Treetagger, Flemm)
- Approche : catégorisation supervisée
- Données de référence :
  - annotation manuelle indépendante par 3 personnes :
    - A1, A2, A3
    - unanimité
    - majorité
  - catégories :
    1. *Je peux comprendre*
    2. *Je ne suis pas sûr*
    3. *Je ne peux pas comprendre*

Détection de mots/passages difficiles

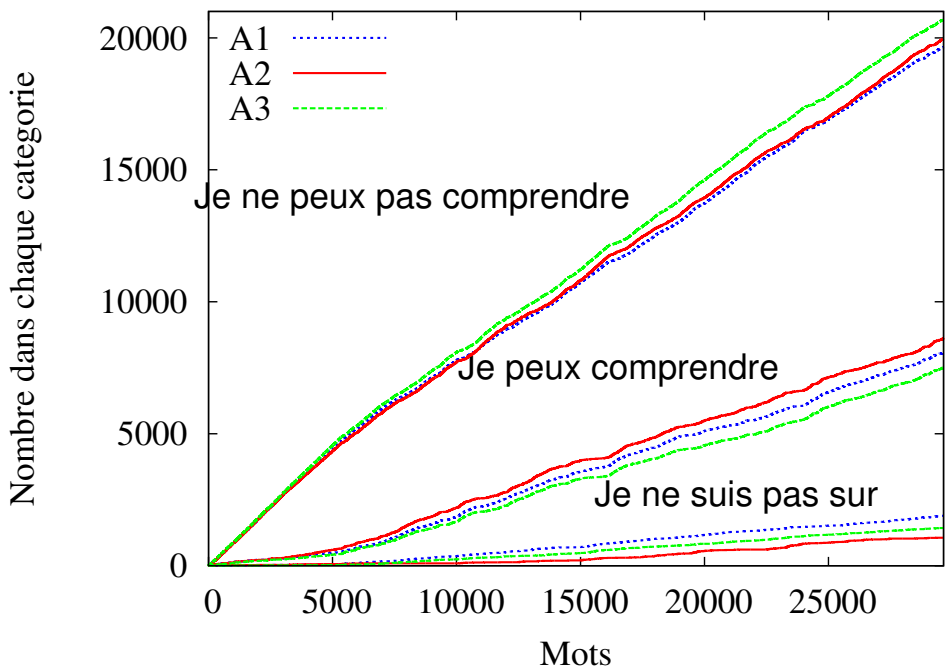
Matériel

- composés (*abdominoplastie*, *dermabrasion*)
- construits (*cardiaque*, *acineux*, *lipoïde*)
- simples (*acné*, *fragment*)

Détection de mots/passages difficiles

Annotation

Cat.	*A1 (%)*	*A2 (%)*	*A3 (%)*	*Unan. (%)*	*Major. (%)*
1.	8 099 (28)	8 625 (29)	7 529 (25)	5 960 (26)	7 655 (27)
2.	1 895 (6)	1 062 (4)	1 431 (5)	61 (0,3)	597 (2)
3.	19 647 (66)	19 954 (67)	20 681 (70)	16 904 (73,7)	20 511 (71)
Total	29 641	29 641	29 641	22 925	28 763



Accord inter-annotateur : Kappa Fleiss 0.735, Kappa Cohen 0.736

Détection de mots/passages difficiles

Descripteurs

24 descripteurs linguistiques et extra-linguistiques :

- *Catégories syntaxiques.* TreeTagger [Schmid, 1994] et Flemm [Namer, 2000] (noms, adjectifs, noms propres, verbes, abréviations) ;
- *Lexiques de référence.* TLFi et lexique.org ;
- *Fréquence sur un moteur de recherche* ;
- *Fréquence dans la terminologie medicale* ;
- *Nombre de types sémantiques* ;
- *Longueur de mots* (nombre de caractères et syllabes) ;
- *Nombre de bases et affixes.* Analyseur morphologique Dérif [Namer, 2003] ;
- *Chaînes initiales et finales.* 3 à 5 caractères ;
- *Nombre et % de consonnes, voyelles et autres caractères* ;
- *Scores de lisibilité classiques.* [Flesch, 1948] et Flesch-Kincaid [Kincaid *et al.*, 1975].

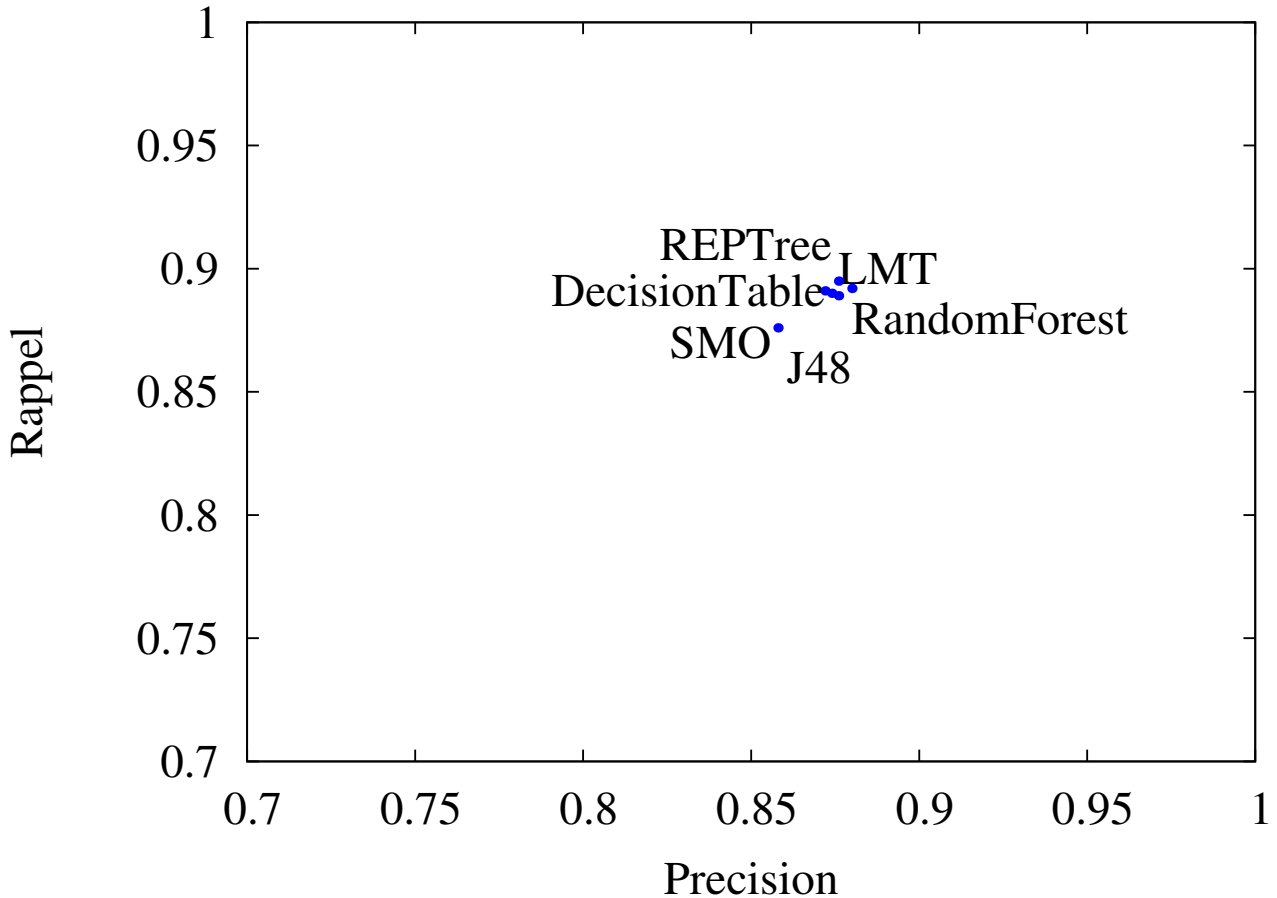
Détection de mots/passages difficiles

Catégorisation supervisée

- Catégorisation avec WEKA
- Cinq ensembles de référence :
  - 3 ensembles : annotations des trois annotateurs (29 641 mots),
  - ensemble *unanimité*, tous les annotateurs sont d'accord (22 925 mots),
  - ensemble *majorité*, accord majoritaire des annotateurs (28 763 mots).
- Distinction entre les mots compréhensibles et non-compréhensibles
- Pertinence des descripteurs
- Baseline : catégorisation dans la catégorie majoritaire

Détection de mots/passages difficiles

Résultats de la catégorisation



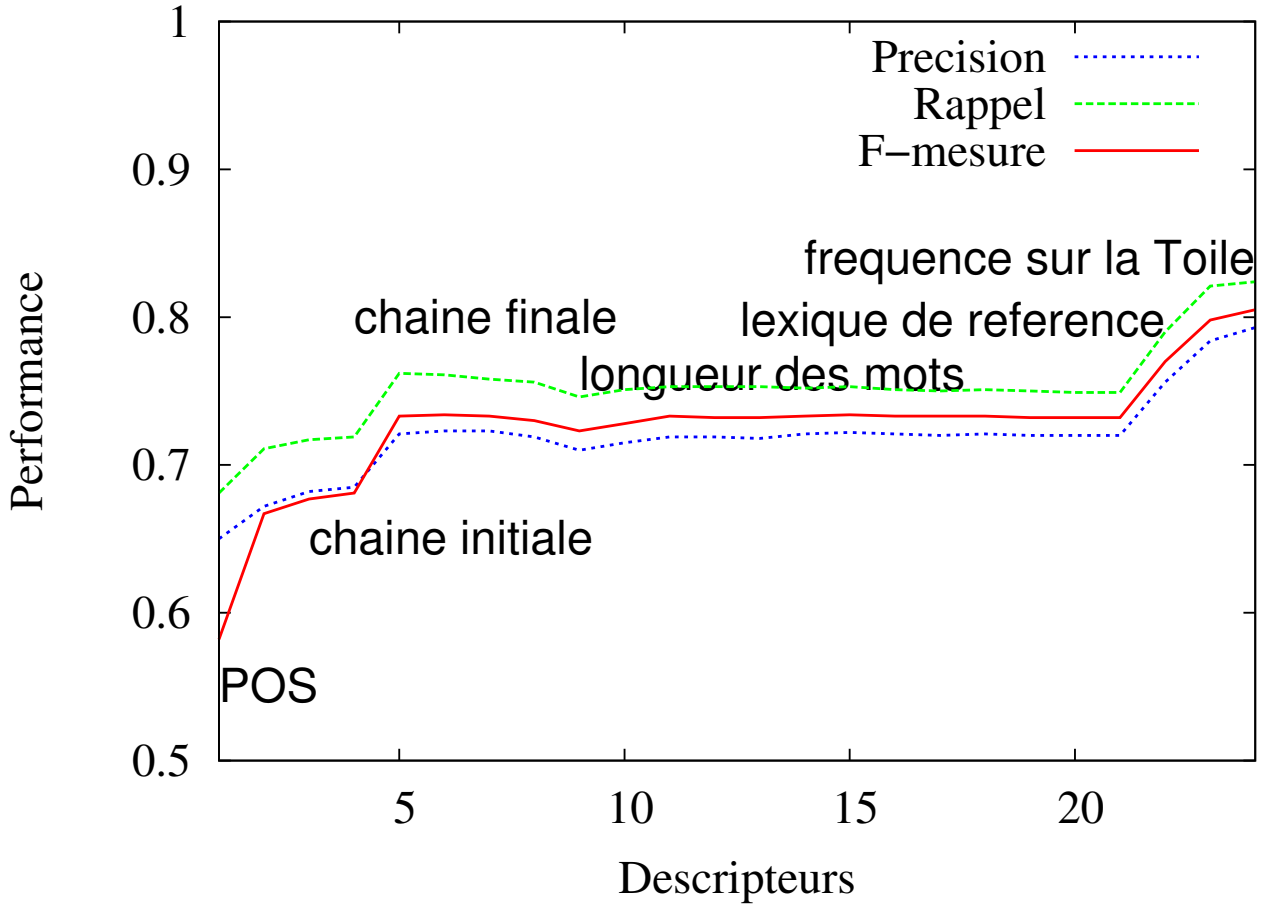
Détection de mots/passages difficiles

Résultats de la catégorisation J48

	A1	*A2*	*A3*	*Una.*	*Maj.*
$\mathcal{P}$	0.794	0.809	0.834	0.946	0.876
$\mathcal{R}$	0.825	0.826	0.862	0.949	0.889
$\mathcal{F}$	0.806	0.814	0.845	0.947	0.881
BL	0.66	0.67	0.70	0.74	0.71
gain	0.14	0.13	0.14	0.20	0.16

Détection de mots/passages difficiles

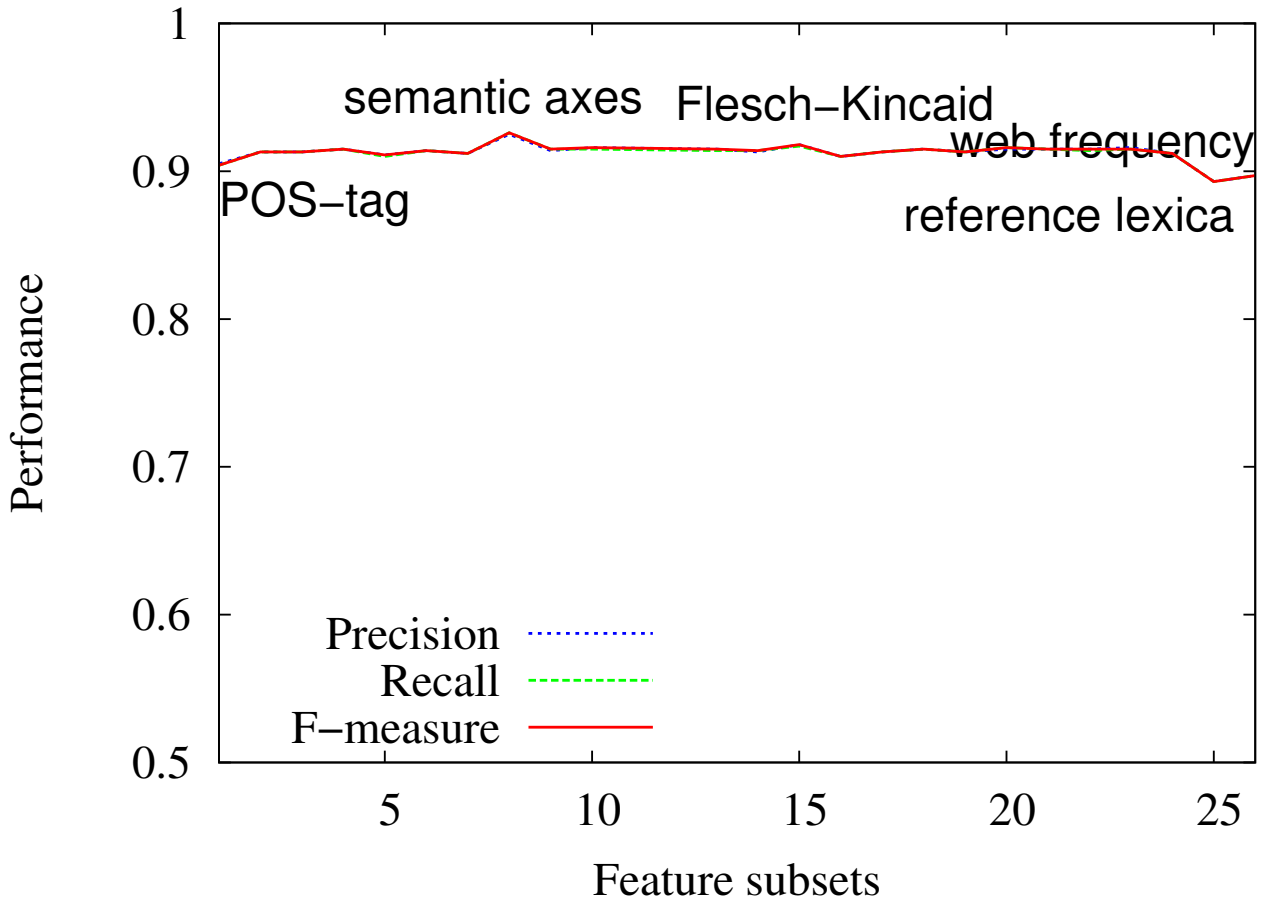
Ajout incrémental des descripteurs



- scores de lisibilité, fréquence dans la terminologie, nombre de types sémantiques

Détection de mots/passages difficiles

Take one out



- types sémantiques, scores de lisibilité

Détection de mots/passages difficiles

Texte brut

Histoire de la maladie
Le patient a été hospitalisé le 18/7/11 à [redacted] pour un AVC ischémique dans le territoire profond de l'artère cérébrale postérieure droite, thrombolysé à H+3.

Le patient présente, comme déficit, une hypoesthésie gauche et une parésie gauche (force motrice à 1/5 au membre supérieur gauche et 2/5 au membre inférieur gauche), un NIHSS à 8, une désorientation tempora-spatiale et une vigilance fluctuante.
Dans les suites, est survenu un OAP post thrombolyse, probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse).

Le patient est donc transféré en réanimation : l'OAP est résolutif sous VNI et oxygénothérapie.

La majoration de l'insuffisance rénale nécessite 2 cures de dialyse. Mr K. est ensuite transféré en post-réanimation devant l'évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge.

Le 11/8/2011, le patient présente une douleur thoracique associée à une désaturation à 83 %, il est donc transféré en Unité de soins intensifs cardiologiques. Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire. Une anticoagulation curative par CALCIPARINE est mise en place.

Détection de mots/passages difficiles

Texte annoté

Histoire de la maladie
Le patient a été hospitalisé le 18 / 7 / 11 à [redacted] pour un AVC ischémique dans le territoire profond de l' artère cérébrale postérieure droite , thrombolysé à H + 3 .

Le patient présente , comme déficit , une hypoesthésie gauche et une parésie gauche (force motrice à 1 / 5 au membre supérieur gauche et 2 / 5 au membre inférieur gauche) , un NIHSS à 8 , une désorientation tempora-spatiale et une vigilance fluctuante .
Dans les suites , est survenu un OAP post thrombolyse , probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse) .

Le patient est donc transféré en réanimation : l' OAP est résolutif sous VNI et oxygénothérapie .

La majoration de l' insuffisance rénale nécessite 2 cures de dialyse . Mr K . est ensuite transféré en post-réanimation devant l' évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge .

Le 11 / 8 / 2011 , le patient présente une douleur thoracique associée à une désaturation à 83 % , il est donc transféré en Unité de soins intensifs cardiologiques . Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire . Une anticoagulation curative par CALCIPARINE est mise en place .

Détection de mots/passages difficiles

Analyse des erreurs et des limites

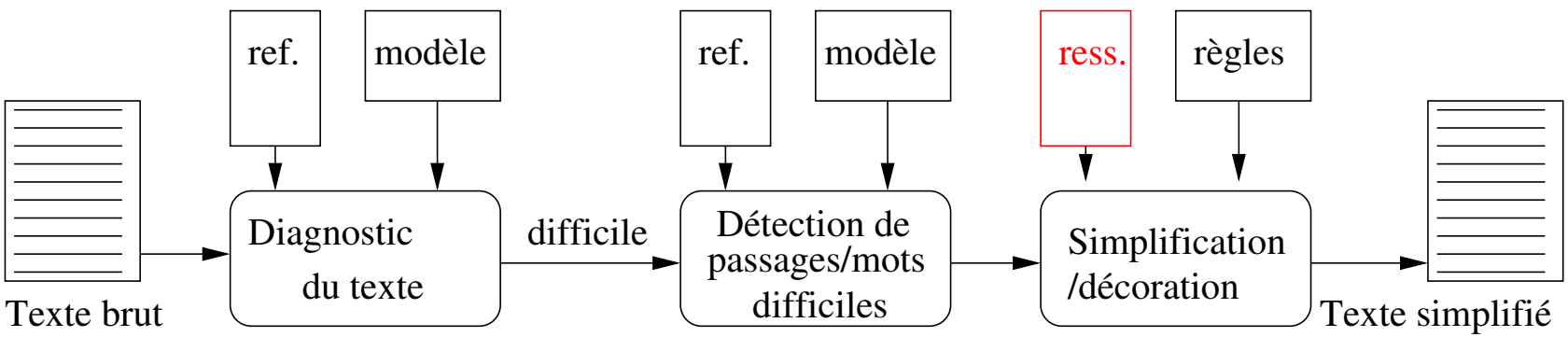
- Entités nommées (*France*, *Indiana*, *Nancy*, *Tokyo*)
  - OK pour les annotateurs, KO pour la catégorisation
- Anatomie humaine (*cloacal*, *pubovaginal*, *nasopharyngé*, *mitral*, *diaphragmatique*, *inguinal*, *strontium*, *érythème*)
  - très souvent : incompréhensibles pour les annotateurs
- Composés (*antihémophile*, *pseudohémophilie*, *sclérodermie*, *hydrolase*, *orthotopique*, *tympanectomie*, *arthrodèse*, *synesthésie*)
  - considérés comme compréhensibles à tord
- Mots avec - (*intestin-côlon*, *semi-fermé*, *post-cataracte*, *non-réponse*, *non-érotique*, *celle-ci*, *sous-rétinien*)
  - considérés comme non compréhensibles à tord
- Fautes d'orthographe (*oreillle*, *épaiseur*)
- Formes fléchies et dérivées
- Entités syntaxiquement complexes (*AVC ischémique*, *embolie pulmonaire basale*, *scintigraphie pulmonaire*)

Détection de mots/passages difficiles

Bilan

- Catégorisation des mots en compréhensible ou non
- Apprentissage supervisé
- Bonnes performances
- Lien avec le texte
- Limites :
  - entités syntaxiquement complexes
  - formes fléchies et dérivées
  - orthographe
  - ...

Acquisition de ressources pour la simplification



Acquisition de ressources pour la simplification

Travaux existants

- Corpus alignés au niveau des phrases [Biran *et al.*, 2011] :
  - In 1900, Omaha was the center of a national uproar over the kidnapping of Edward Cudahy, Jr., the son of a local meatpacking **magnate**.
  - OUTPUT : In 1900, Omaha was the center of a national uproar over the kidnapping of Edward Cudahy, Jr., the son of a local meatpacking **businessman**.
  - {*magnate*, *king*}, {*magnate*, *businessman*}

Acquisition de ressources pour la simplification

Motivation

- Besoin d'avoir des ressources dédiées
- " Traduire " les termes difficiles
- Souvent, des glossaires {*difficile*, *facile*}
  - [Zeng *et al.*, 2006] :
    {*myocardial infarction*, *heart attack*}
    {*abortion*, *termination of pregnancy*}
    {*acrodynia*, *pink disease*}
  - [Deléger & Zweigenbaum, 2008] :
    {*consommation régulière*, *consommer de façon régulière*}
    {*gêne à la lecture*, *empêche de lire*}
    {*évolution de l'affection*, *la maladie évolue*}
  - [Cartoni & Deléger, 2011] :
    {*retard de cicatrisation*, *retarder la cicatrisation*}
    {*apports caloriques*, *apport en calories*}
    {*calculer les doses*, *doses sont calculées*}
    {*efficacité est renforcée*, *renforcer son efficacité*}

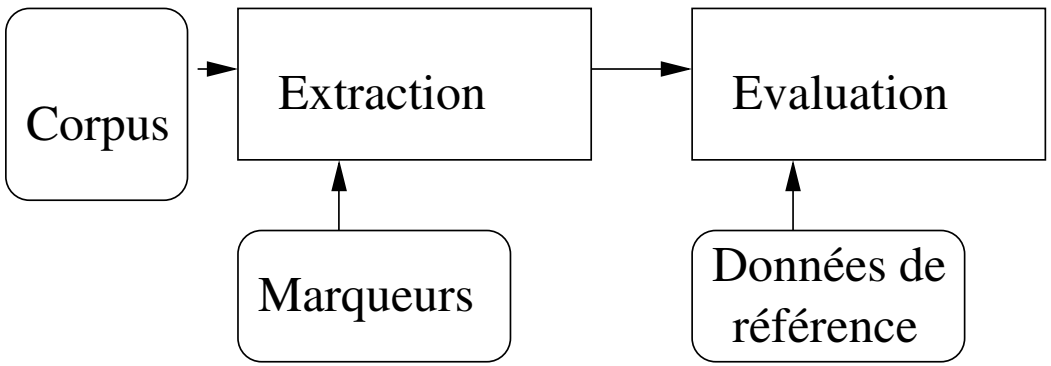
Acquisition de ressources pour la simplification

Expériences

1. Contextes définitoires

   [Grabar & Hamon, 2016]
2. Compositionnalité morphologique des termes
3. Reformulations

Contextes définitoires



Contextes définitoires

Matériel

- Termes :
  - Snomed International [Côté *et al.*, 1993], partie française d'UMLS [Lindberg *et al.*, 1993]
    - mots des termes
    - pas de nombres
- Corpus :
  - Wikipédia, Portail de la Médecine
    - version de janvier 2015
    - 18 434 articles
    - 15 235 219 occurrences

Contextes définitoires

Méthode

- Définition : structure avec deux éléments :
  - *definiendum* (terme à définir) et *definiens* (la définition)
  - *Myocarde est le tissu musculaire du coeur*
- Application de quatre patrons [Péry-Woodley & Rebeyrolle, 1998]
  - *désigne*
  - *est un*
  - *est appelé*
  - *peut être défini comme*
- ...avec des variations flexionnelles
- Déclencheur : terme

Contextes définitoires

Résultats

- Extraction :
  - 2 037 contextes définitoires
  - 1 286 termes uniques
- Type de termes définis :
  - composés :
    hypoglycémie, *acidocétose*, *angiographie*, *hypokaliémie*,
  - mots affixés :
    curetage, *capsulite*, *arthrose*, *glaucome*, *durillon*, *pré-diabète*,
  - mots morphologiquement non construits :
    cataracte, *impétigo*, *zona*

Contextes définitoires

Résultats

Définitions correctes :

- *L'hypoglycémie est un manque de sucre dans l'organisme*
- *Une septicémie est un empoisonnement du sang du à un microbe*
- *Le curetage est un nettoyage en profondeur d'une gencive inflammée*
- *Pour un être humain adulte, une hypoglycémie est une glycémie inférieure à 0,8 g/L*
- *Les signes classiques annonciateurs de l'hypoglycémie sont des sueurs, pâleur, palpitations, fringales en particulier*
- *L'impétigo est une infection cutanée, qui provoque des pustules qui dégénèrent en croûtes jaunâtres, l'impétigo est due à...*

Contextes définitoires

Résultats

Définitions possiblement correctes :

- *La mélancolie est une douceur qui nous berce*
- *Une injection est une agression, qui sauve, mais c'est quand même une agression*

Contextes définitoires

Résultats

- Compréhension (*péricarde*) :
  - \+ *La couche extérieure du cœur est appelée péricarde.*
  - ∼ *Le péricarde est un sac à double paroi contenant le cœur et les racines des gros vaisseaux sanguins.*
  - − *Le péricarde est un organe de glissement, formé de deux feuillets limitant une cavité virtuelle, la cavité péricardique, qui permet les mouvements cardiaques.*

Contextes définitoires

Résultats

- Évaluation :
  - précision stricte : 52,5 %
    - définitions correctes : 849
  - précision lâche : 68 %
    - définitions correctes et possiblement correctes : 1 028

Contextes définitoires

Bilan

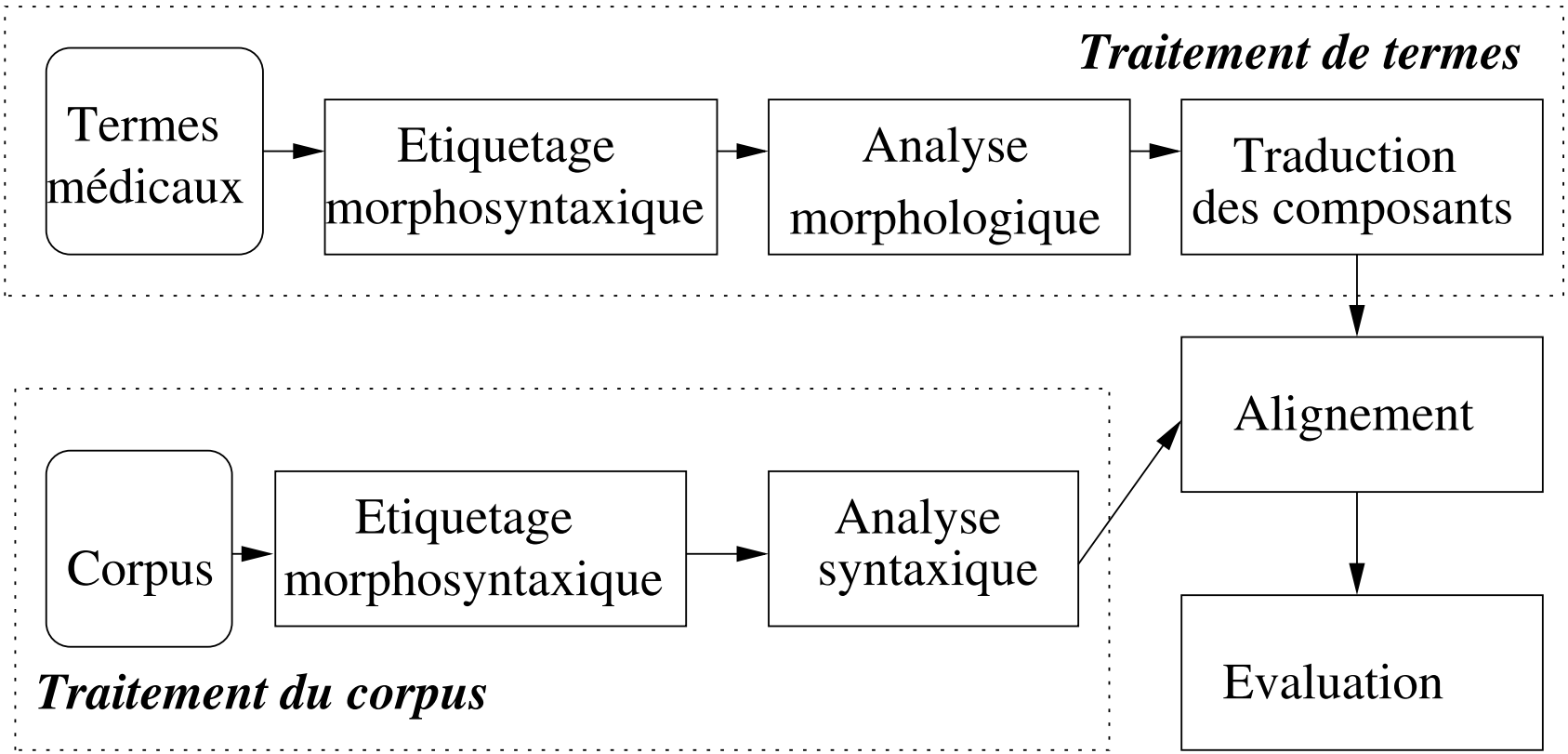
- Acquisition de définitions de termes médicaux
- Différents types de termes
  - non construits, affixés, composés néoclassiques
- Résultats :
  - jusqu'à 1 028 termes
- Précision :
  - stricte : 52,5 %
  - lâche : 68 %

Acquisition de ressources pour la simplification

Expériences

1. Contextes définitoires
2. Compositionnalité morphologique des termes
   [Grabar & Hamon, 2016]
3. Reformulations

Composition morphologique



Composition morphologique

Matériel

- Termes :
  - Snomed International [Côté *et al.*, 1993], partie française d'UMLS [Lindberg *et al.*, 1993]
    - mots des termes
    - pas de nombres
- Corpus :
  - Wikipédia, Portail de la Médecine
    - version de janvier 2015
    - 18 434 articles
    - 15 235 219 occurrences
- Ressources linguistiques :
  - liste de mots de vides
  - morphologique : 163 823 paires de mots (dérivations, flexions)

Composition morphologique

1. Traitement de termes médicaux

- Étiquetage morpho-syntaxique et lemmatisation `Cordial` [Laurent *et al.*, 2009]
  - *myocardique/A*, *cholécystectomie/N*
- Analyse morphologique `DériF` [Namer, 2003]
  - *myocardique/A* : *[[[myo N*] [carde N*] NOM] ique ADJ]*
  - *cholécystectomie/N* : *[[cholécysto N*] [ectomie N*] NOM]*
- Association avec les mots du français (ressource supplétive)
  - *myocardique/A* :
    - *myo*=*muscle*, *carde*=*cœur*
  - *cholécystectomie/N* :
    - *cholécysto*=*vésicule biliaire*, *ectomie*=*ablation*

Composition morphologique

2. Traitement du corpus

- `Cordial` [Laurent *et al.*, 2009]
  - étiquetage morpho-syntaxique et lemmatisation
  - analyse syntaxique
- Définir les frontières des syntagmes

Composition morphologique

3. Extraction de paraphrases

- Mise en parallèle :
  - syntagmes et décompositions morphologiques des termes
- Tout type de contextes :
  - *Les causes de tachycardie ventriculaire sont superposables à celles des extrasystoles ventriculaires : infarctus du myocarde, insuffisance cardiaque, hypertrophie du muscle du cœur et prolapsus de la valve mitrale.*

  $\Rightarrow$ {*myocarde*, *muscle du cœur*}
- Quatre paramètres à varier :
  1. taille de la fenêtre : 1, 2, 3 syntagmes
  2. ressources linguistiques :
     - formes brutes
     - ressources morphologiques (flexions, dérivations)
     - ressource de synonymes
  3. taux d'alignement des termes
  4. taux d'alignement des syntagmes

Composition morphologique

4. Évaluation

- Validation :
  1. paraphrase correcte : {*myocardique*, *muscle du cœur*},
  2. analyse morphologique incorrecte : {*sanglot*, *lot sang*}
  3. traduction vers le français incorrecte : *antisolaire*, {*sol*, *sol*} au lieu de {*sol*, *solaire*}
  4. informations correctes au milieu d'autres informations, informations partielles
     - partiel : {*endophtalmie*, *interne de l'œil*}
     - complet : *inflammation des tissus internes de l'œil*
  5. extraction fausse
- Précision :
  - précision stricte $P_{stricte}$: cas 1
  - précision lâche P_{lache} : cas 1 et 4
  - taux d'erreurs : cas 5
  - cas 2 et 3 : pas pris en compte

Composition morphologique

Résultats

- 274 131 termes UMLS et Snomed International
- 76 536 mots sans nombres
- 15 121 mots analysés par `Dérif`
  - décomposés en deux bases au moins
- Alignement syntagme/terme (pourcentage d'alignement) :

  *E*1 : terme et syntagme complets dans l'alignement :
  - {*myo pathie*, *maladie du muscle*}

  *E*2 : terme complet, syntagme partiel :
  - {*myo pathie*, *maladie du muscle cardiaque*}

  *E*3 : terme partiel, syntagme complet :
  - {*myopathie*, *la maladie*}

  *E*4 : terme et syntagme partiels :
  - {*myopathie*, *l' origine de la maladie*}
- Travail avec E1 (le plus optimisé)

Composition morphologique

Extraction de paraphrases

Nb de	*unigrammes*			*bigrammes*			*trigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
syntagme	9854	16093	22110	11875	18504	27670	7936	12284	19984
terme unique	1513	1947	2090	1780	2260	2463	1523	1966	2231
$syntagme_{E1}$	2681	4163	5370	1109	1611	2521	403	634	988
$terme\ unique_{E1}$	668	1023	1051	492	670	962	239	358	472

- total et E1
- ressources linguistiques : augmentent le volume
  - *b* : sans les ressources
  - *l* : ressources morphologiques
  - *s* : ressources de synonymie
- n-grammes de syntagmes : diminuent le volume
  - seuil d'alignement acceptable

Composition morphologique

Évaluation

Nombre de	*unigrammes*			*bigrammes*			*trigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
paraphrases correctes	549	785	644	378	517	461	195	290	257
possibl. correctes	39	32	67	22	45	75	10	19	41
traitement de termes	47	60	44	28	28	46	9	10	26
paraphrase incorrectes	33	146	296	64	80	380	25	39	148
$P_{stricte}$	82	77	61	77	77	48	82	81	55
P_{lache}	88	80	68	81	84	40	86	86	63
$\%_{incorrect}$	5	14	28	13	12	39	11	11	31

- Évaluation :
  - précision stricte 82 à 55 %
  - précision lâche 86 à 40 %
  - taux d'erreurs 5 à 39 %
- Ressources
  - sans ressources : précision la plus élevée
  - ressources morphologique : bonne précision
  - ressources de synonymie : la plus faible précision

Composition morphologique

Analyse morphologique

- Analyse ambiguë
  - *[post [[uro N*] [graphie N*] NOM] NOM]*
  - *[[posturo N*] [graphie N*] NOM]*
- Analyse incorrecte
  - *sanglot* : *lot* et *sang*
  - *exotique* : *externe* et *oreille*

Composition morphologique

Extraction de paraphrases et leur évaluation

Extraction de paraphrases correctes

- Brut
  - *podalgie : douleur du pied*
  - *mastite : inflammation du sein*
  - *cystoprostatectomie : ablation de la vessie et de la prostate*
- Morphologie
  - *desmorrhexie* : *rupture des ligaments* (ligament→ligaments)
  - *bronchite : inflammation des bronches, inflammation bronchique* (bronche→bronches, bronche→bronchique)
  - *dentalgie : douleurs dentaires* (dents→dentaires)
- Synonymie
  - *aclasie : absence de fracture* (cassure→fracture)
  - *enterectomie : résection des intestins* (ablation→résection)

Composition morphologique

Extraction de paraphrases et leur évaluation

- Relations sémantiques entre composants :
  - bien gérées sur la base du corpus
  - erreurs : coordination/subordination
    - *hematospermie* : *le sang ou le sperme*, au lieu de

    → *le sang dans le sperme*
- Termes non compositionnels :
  - *ostéodermie* : *peau* et *os*, au lieu de

    → *une structure d'écailles, de plaques osseuses ou d'autres compositions dans les couches dermiques de la peau, comme chez les lézards ou dinosaures*
- Couverture des 15 121 termes analysés morphologiquement :
  - 6,8 % (1 031) paraphrases correctes
  - 7,5 % (1 128) paraphrases correctes et possiblement correctes correctes

Composition morphologique

Ressources linguistiques

Synonymie : valeurs sémantiques contextuelles

Peut extraire des paraphrases incorrectes :

- *cardialgie* :
  - correct : *douleur de cœur*
  - extrait : *plaie du cœur* (douleur→plaie)
- *cheiropathie* :
  - correct : *maladie des mains*
  - extrait : *Le syndrome main* (maladie→syndrome)
- *cinépathie*
  - correct : *mal des transports*
  - décomposé en *mouvement* et *maladie*
  - extrait : *évolution du syndrome* (mouvement→évolution, maladie→syndrome)

Composition morphologique

Termes non paraphrasés

- Plus de 2 composants :
  - *hémi-desmo-some*, *hémo-histio-blaste*
- Composants et leurs combinaisons rares :
  - *hémi-desmo-some* : *demi*, *ligament*, *corpuscule*
- Ressource supplétive :
  - trop stricte
  - d'autres méthodes [Claveau & Kijak, 2014]

Composition morphologique

Bilan

- Paraphrases grand public pour les termes médicaux
- Composés néoclassiques
- Résultats :
  - jusqu'à 1 128 termes
- Précision moyenne :
  - toutes les expériences : 76 %
  - sans synonymes : 86 %

Acquisition de ressources pour la simplification

1. Contextes définitoires
2. Compositionnalité morphologique des termes
3. Reformulations

   [Antoine & Grabar, 2016]

Reformulations

Hypothèse

- Paraphrase : un même concept exprimé avec des moyens linguistiques différents :
  - *Google a acheté Youtube* $\rightarrow$ *Youtube a été vendu à Google*
- Reformulation : redire différemment ce qui a déjà été dit [Le Bot *et al.*, 2008]
- Présence de reformulations :
  - indique les mots/termes difficiles
  - offre les indices pour l'extraction

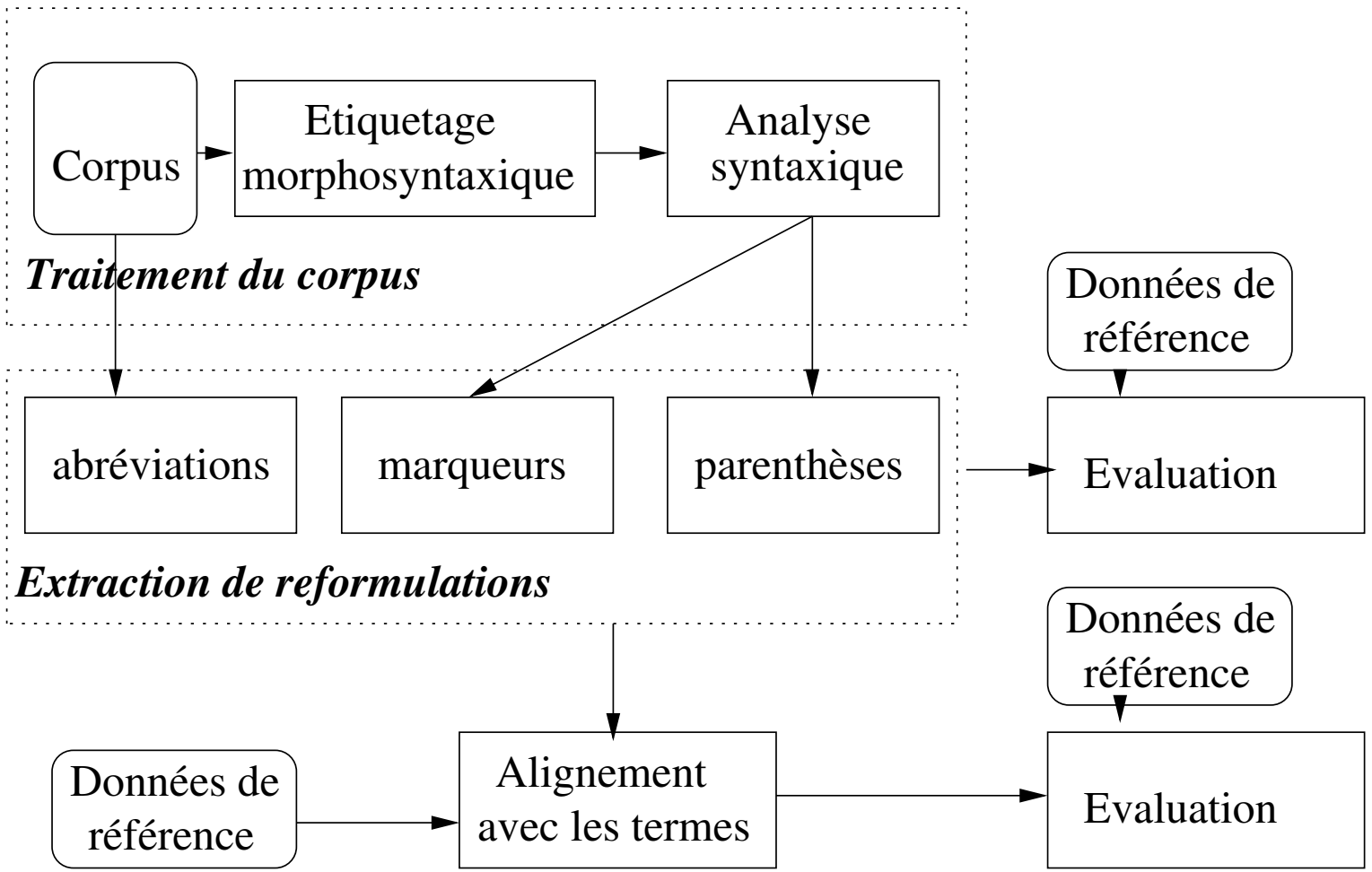
Reformulations

Corpus et ressources

- Corpus : monolingues simples, discours médical
  - développement : forum *masante.net*
    - 6 139 réponses, 315 362 occurrences
      Cher(e) pseudonyme, réponse du médecin. *Bien cordialement. Ceci n'est pas une consultation médicale et n'a pas pour objet de la remplacer.*
  - test : Wikipédia, Portail de la Médecine
    - 18 434 articles, 15 235 219 occurrences
- Ressources linguistiques :
  - liste de mots de vides
  - morphologique : 163 823 paires de mots (dérivations, flexions)
- Terminologie médicale en français :
  - UMLS : Unified Medical Language System
    [Lindberg *et al.*, 1993]
  - SNOMED Int : Systematized Nomenclature of Medicine
    [Côté *et al.*, 1993]

Reformulations

Schéma général de l'approche



Reformulations

Extraction de siglaisons et de leurs formes étendues

- Inspiré de [Schwartz & Hearst, 2003]
- 2 types de patrons :
  1. *anti-inflammatoires non stéroïdiens (AINS)*
  2. *AVC (Accident Vasculaire Cérébral)*
- Utilisation du texte brut
- Reconnaissance : majuscules, parenthèses
- Association lettre $\rightarrow$ mot
- Gestion des doublons : *leucémie aiguë lymphoblastique (LAL)*

Reformulations

Extraction de reformulations avec marqueurs

concept marqueur reformulation
vésiculaire, c'est-à-dire, venant de la vésicule biliaire

- 3 marqueurs :
  - *c'est-à-dire*
  - *autrement dit ; Autrement dit*
  - *encore appelé(e)(s)*
- Pré-traitement
- Étiquetage et analyse morpho-syntaxique de Cordial [Laurent *et al.*, 2009]
- Déclencheur : marqueurs
- Récupération du concept et de la reformulation :
  - informations syntaxiques

Reformulations

Extraction de reformulations avec marqueurs

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
Vous	vous	PPER2P	Pp2.pn	1	S	1
ne	ne	ADV	Rpn	3—1	S	1
devez	devoir	VINDP2P	Vmip2p	3	V	1
pas	pas	ADV	Rgn	3	Q	1
employer	employer	VINF Vmn	–	5	D	2
de	de	PREP	Sp	7	D	2
savons	savon	NCMP	Ncmp	7	D	2
ou	ou	COO	Cc	7	F	2
des	de le	DETDPIG	Da-.p-i	10—7	F	2
laits	lait	NCMP	Ncmp	10—7	F	2
sophistiqués	sophistiqué	ADJMP	Afpmp	10—7	F	2
,	,	PCTFAIB	Ypw	-	-	2
c'	ce	PDS	Pd-..-	13	N	2
est	est	ADV	Rgp	-	p	2
-à	à	PREP	Sp	16	F	2
-dire	dire	VINF	Vmn–	16	F	2
contenant	contenant	NCMS	Ncms	17	D	2
plusieurs	plusieurs	ADJIND	Dt-.p-	19	D	2
composants	composant	NCMP	Ncmp	19	D	2

Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire, contenant plusieurs composants

Reformulations

Extraction de reformulations avec marqueurs

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
Vous	vous	PPER2P	Pp2.pn	1	S	1
ne	ne	ADV	Rpn	3—1	S	1
devez	devoir	VINDP2P	Vmip2p	3	V	1
pas	pas	ADV	Rgn	3	Q	1
employer	employer	VINF Vmn	–	5	D	2
de	de	PREP	Sp	7	D	2
savons	savon	NCMP	Ncmp	7	D	2
ou	ou	COO	Cc	7	F	2
des	de le	DETDPIG	Da-.p-i	10—7	F	2
laits	lait	NCMP	Ncmp	10—7	F	2
sophistiqués	sophistiqué	ADJMP	Afpmp	10—7	F	2
,	,	PCTFAIB	Ypw	-	-	2
c'	ce	PDS	Pd-..-	13	N	2
est	est	ADV	Rgp	-	p	2
-à	à	PREP	Sp	16	F	2
-dire	dire	VINF	Vmn–	16	F	2
contenant	contenant	NCMS	Ncms	17	D	2
plusieurs	plusieurs	ADJIND	Dt-.p-	19	D	2
composants	composant	NCMP	Ncmp	19	D	2

Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire, contenant plusieurs composants

Reformulations

Extraction de reformulations avec parenthèses

concept (reformulation)
avec des prélèvements (biopsie)

- Pré-traitement
- Étiquetage et analyse morpho-syntaxique de Cordial [Laurent *et al.*, 2009]
- Déclencheur : parenthèses
- Filtres pour limiter le bruit :
  - *un problème hormonal (thyroïde, surrénale)*
- Extraction :
  - concept : informations syntaxiques
  - reformulation : entre parenthèses

Reformulations

Extraction de reformulations avec parenthèses

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
une	un	DETIFS	Da-fs-i	13	N	2
gastroscopie	gastroscopie	NCFS	Ncfs	13	N	2
avec	avec	PREP	Sp	16	H	2
des	de le	DETDPIG	Da-.p-i	16	H	2
prélèvements	prélèvement	NCMP	Ncmp	16	H	2
(	(	PCTFAIB	Ypo	-	-	2
biopsie	biopsie	NCFS	Ncfs	18	N	2
)	)	PCTFAIB	Ypc	-	-	2
.	.	PCTFORTE	Yps	-	-	-

une gastroscopie avec des prélèvements (biopsie)

Reformulations

Évaluation des extractions

- Préparation des données de référence des extractions
- Toutes les phrases avec les reformulations
- Annotations de reformulations avec un guide d'annotation
  - *<C>d'origine labyrinthique</C>, <M>c'est à dire</M>, <Rgen>venant de l'oreille interne</Rgen>*
- Accord inter-annotateur : kappa de Cohen [Cohen, 1960]
  - 2 niveaux : phrase et token
  - accord binaire : O/N

	Extraction	
	Phrase	*Token*
Abréviations	0,661	0,967
Marqueurs	0,24	0,816
Parenthèses	0,651	0,575

Reformulations

Alignement avec une terminologie médicale

- Comparaison des segments extraits avec les termes de la terminologie médicale :
  - évite les extractions non pertinentes :
    - *en fibres (pas trop vite sinon vous serez ballonnée)*
  - fait ressortir les segments pertinents et exploitables
- Méthode :
  - casse, désaccentuation, normalisation morphologique
  - suppression des mots vides
  - choix du taux d'alignement : segments, termes

Reformulations

Évaluation de l'alignement

- Données de référence :
  - à partir de l'alignement aux seuils 40/40, corpus de développement
  - deux annotateurs, consensus
- Mesure d'évaluation : précision
- Définition des seuils sur le corpus de développement
- Application de ces seuils sur le corpus de test

	Alignement
Abréviations	0,208
Marqueurs	0,714
Parenthèses	0,817

Reformulations

Résultats : Extractions des siglaisons

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- Types d'extractions :
  - Complètes : *AINS : anti inflammatoire non stéroïdien*
  - Partielles mais correctes : *CIV : communication interventriculaire*
  - Partielles et exploitables : *CHU : hôpital universitaire*
  - Partielles et inexploitables : *NFS : faire sang*
  - Pas d'extraction : *comment sont les ALAT(ou SGPT) et les ASAT (ou SGOT)*

Résultats

Extractions des reformulations avec marqueurs

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- Trois marqueurs :
  - *c'est-à-dire* : 80/1 929
  - *a-Autrement dit* : 8/145
  - *encore appelé(e)(s)* : 8/86
- Difficultés : détection de frontières
  - *une toxi-infection, c'est-à-dire, qu' elle peut*
  - *Une salpingite, c'est-à-dire, une inflammation des trompes est possible*
  - *en, c'est-à-dire, au contact du sang circulant*
  - *des dilations des canaux galactophores, c'est-à-dire qui fabriquent le lait*

Résultats

Extractions des reformulations avec parenthèses

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- Difficultés :
  - frontière des concepts :
    - *une greffe de valve prothétique (valve mécanique artificielle)*
    - *se bouche (hémorroïdes)*
  - extractions non pertinentes :
    - *énergétique (carence plutôt liée au marasme)*

Évaluation des extractions

Précision, rappel et F-mesure des extractions pour chaque méthode

	Abréviations			*Marqueurs*			*Parenthèses*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
exact	0.74	0.74	0.74	0.24	0.24	0.24	0.23	0.23	0.23
inexact	0.94	0.94	0.94	0.98	0.98	0.98	0.68	0.68	0.68

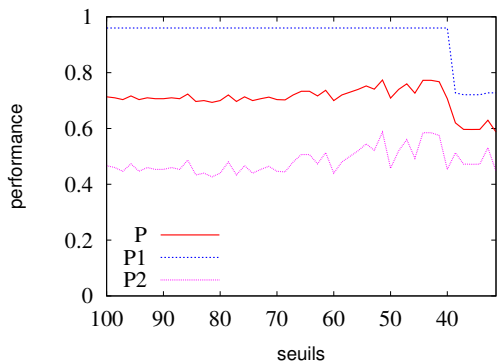
- corpus de développement
- données de référence consensuelles
- script d'évaluation : DEFT 2015, tâche 3
- Bilan :
  - fiabilité des extractions de siglaisons
  - pertinence des reformulations avec marqueurs
  - bruit des reformulations avec parenthèses
  - recouvrement entre marqueurs et parenthèses : 0,007%

Alignement avec la terminologie

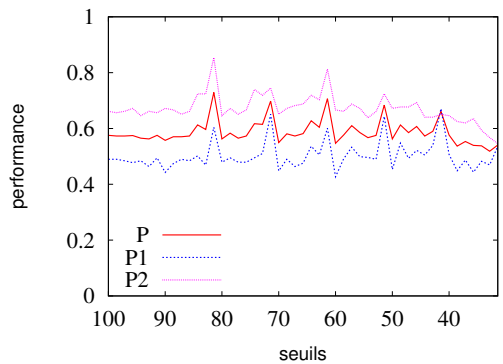
- Types d'alignement de termes :
  - proposition pertinente :
    - *communication interventriculaire : communication interventriculaire.C0018818...*
  - proposition avec variation morpho-syntaxique :
    - *troubles gastrointestinaux fonctionnels/C0559031.T047.DISO*
    - *troubles gastro intestinaux fonctionnels/C0559031.T047.DISO*
  - proposition partielle :
    - *semaines amenorrhee : amenorrhee/C0002453.T047.DISO*
  - proposition compositionnelle (*cause de pus*) :
    - *cause/C0085978.T078.CONC/...*
    - *pus/C0034161.T031.ANAT/...*
  - Proposition non pertinente :
    - *LCR : ph lcr/C0853364*
    - *liquide cerebro : regime liquide/C-F2300*
  - Aucune proposition : *NFS : —*

Alignement avec la terminologie

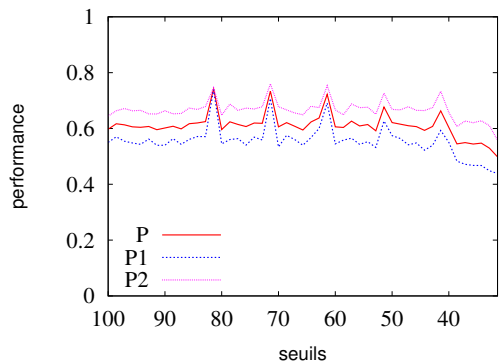
définition des seuils des alignements



(a) Abréviations



(b) Marqueurs



(c) Parenthèses

- corpus de développement
- données de référence consensuelles
- précision des segments alignés :
  - par segment, moyennes

Alignement avec la terminologie

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occurrences	75	96	312	88 762	2 757	100 103
total	11	5	38	154	42	3 738
partiel	44	37	123	1 634	557	25 708
non alignés	20	54	150	6 318	1 937	60 928

- deux segments alignés :
  - *d'une fibromyalgie : fibromyalgie.C0016053.T047.DISO*
  - *SPID (syndrome polyalgique idiopathique diffus) : syndrome polyalgique idiopathique diffus/C0016053.T047.DISO*
- un seul segment aligné :
  - *TSH : –*
  - *thyroïde : thyroide.C0040132.T023.ANAT*
- aucun segment aligné :
  - *HAS : –/Haute Autorité Santé : –*

Typologie des reformulations

- Typologie de l'état de l'art [Bhagat & Hovy, 2013]
- Difficulté de classifier avant les alignements (trop de bruit)
- Reformulations avec marqueurs :
  - synonyme :
    - *l'interruption naturelle ou accidentelle de la grossesse, c'est-à-dire, un avortement spontané*
  - définition :
    - *la contractilité myocardique, c'est-à-dire, la capacité des cellules musculaires myocardiques à se contracter en réponse à un potentiel d'action*
- Reformulations avec parenthèses :
  - synonyme :
    - *nerveux (hystérie)*
  - définition :
    - *une scoliose (courbure de la colonne vertébrale)*
  - relation cause à effet :
    - *d'ulcère tropical (moisissures de la jungle)*

Reformulations

Discussion

- Exploitation de reformulation pour l'acquisition du vocabulaire
- 3 méthodes :
  - abréviation : inspiré de l'algorithme proposé par (Bhagat et al, 2013)
  - marqueurs, parenthèses : observations des données
- Alignement avec une terminologie
- Résultats :
  - meilleurs résultats avec les abréviations (74, 94%)
  - bonne couverture avec les parenthèses
  - bonne pertinence avec les marqueurs
  - taux d'alignement : 65% - 313 (dev) ; 17% - 31 833 (test)

Reformulations

Discussion

- Reformulations dans les corpus à destination du grand public
  - réponses des médecins dans les forums de discussion
  - Wikipédia
- Extraction de segments
  - différents types de segments
- Complémentarité de méthodes
- Alignement avec la terminologie médicale

Acquisition de ressources pour la simplification

Bilan

Comparaison entre les approches

	type terme	nb. para	précision
définitions	tout type	1 028	0,52, 0,68
morphologie	composés	1 128	0,76, 0,86
abréviations	abréviations	42, 8 106	0,74/0,94
marqueurs	tout type	96, 2 710	0,24/0,98
parenthèses	tout type	305, 92 971	0,23/0,68

- propositions souvent différentes
- faible recouvrement
- lien avec les terminologies

Acquisition de ressources pour la simplification

Bilan

Comparaison avec les travaux existants

	type terme	nb. para	précision
[Zeng *et al.*, 2006]	tous	CHV	
[Elhadad & Sutaria, 2007]	tous	152	0,58
[Deléger & Zweigenbaum, 2008]	m-synt.	65, 82	0,67, 0,60
[Cartoni & Deléger, 2011]	m-synt.	109	0,66
définitions	tout type	1 028	0,52, 0,68
morphologie	composés	1 128	0,76, 0,86
abréviations	abréviations	42, 8 106	0,74/0,94
marqueurs	tout type	96, 2 710	0,24/0,98
parenthèses	tout type	305, 92 971	0,23/0,68

- morpho-syntaxique :
  - {*consommation régulière*, *consommer de façon régulière*}
- performances comparables, meilleure couverture
- lien avec les terminologies

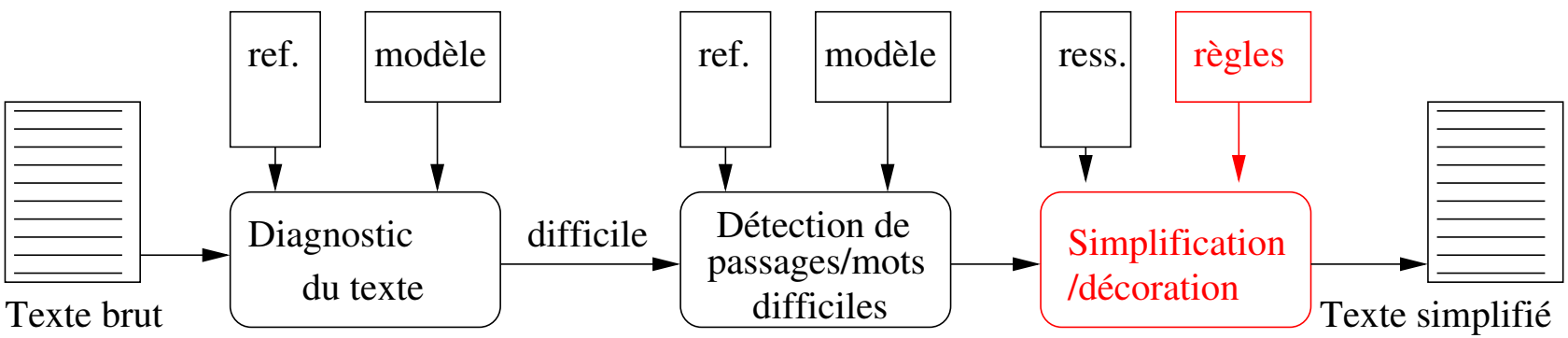
Acquisition de ressources pour la simplification

Bilan

Comparaison avec les travaux existants

- `DériF` [Namer, 2003] :
  - glose en langage artificiel pour tout terme analysé
  - notre méthode : la couverture dépend du contenu des corpus
- *myocarde* :
  - *"(Partie de – Type particulier de) coeur en rapport avec le(s) muscle"*
  - *muscle du coeur*
- *desmorrhexie* :
  - *"rupture (du – liée au) ligament"*
  - *rupture des ligaments*

Simplication/décoration de textes



Simplication/décoration de textes

Motivation

Objectifs :

- Rendre les documents plus facilement compréhensibles :
  - syntaxe
  - sémantique
  - lexique
  - pragmatique, structure de textes

Simplication/décoration de textes

Destinaitaires

- Enfants
  [Son *et al.*, 1008, De Belder & Moens, 2010, Vu *et al.*, 2014],
- Personnes non ou mal-alphabétisées, locuteurs étrangers
  [Paetzold & Specia, 2016],
- Personnes handicapées ou ayant des pathologies neurodégénératives [Chen *et al.*, 2016],
- Personnes non spécialistes face à des documents spécialisés
  [Arya *et al.*, 2011, Leroy *et al.*, 2013]

Simplication/décoration de textes

Applications

Améliorer les résultats de :

- l'analyse syntaxique
  [Chandrasekar & Srinivas, 1997, Jonnalagadda *et al.*, 2009],
- l'annotation sémantique [Vickrey & Koller, 2008],
- le résumé automatique [Blake *et al.*, 2007],
- la traduction automatique
  [Stymne *et al.*, 2013, Štajner & Popović, 2016],
- l'indexation [Wei *et al.*, 2014],
- la recherche et extraction d'information
  [Beigman Klebanov *et al.*, 2004]

Simplication/décoration de textes

Simplification syntaxique

- Objectif :
  - rendre la structure syntaxique des phrases plus légère
- Focalisation sur :
  - subordonnés
  - passifs
- Travail avec les arbres syntaxiques
  [Chandrasekar & Srinivas, 1997, Siddharthan, 2006, Max, 2008, Candido, Jr. *et al.*, 2009, Brouwers *et al.*, 2014].

Simplication/décoration de textes

Simplification syntaxique

- *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville, soit 20 % de la population totale du pays.*
- *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville.*

Simplication/décoration de textes

Simplification syntaxique

- *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville, soit 20 % de la population totale du pays.*
- *Elle compte, selon les autorités du pays, en 2009, 5 878 609 habitants pour l'agglomération, et 3 796 677 habitants pour la ville.*

Simplication/décoration de textes

Simplification syntaxique

- *Les mélodies sont accrocheuses et les arrangements très soignés ; c'est ainsi que "Mamma Mia" et "Fernando" (malgré quelques erreurs de grammaire anglaise) occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.*
- *Les mélodies sont accrocheuses et les arrangements très soignés.*
  C'est ainsi que "Mamma Mia" et "Fernando" occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.

Simplication/décoration de textes

Simplification syntaxique

- *Les mélodies sont accrocheuses et les arrangements très soignés ; c'est ainsi que "Mamma Mia" et "Fernando" (malgré quelques erreurs de grammaire anglaise) occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.*
- *Les mélodies sont accrocheuses et les arrangements très soignés.*
  C'est ainsi que "Mamma Mia" et "Fernando" occupèrent la première place des palmarès mondiaux dans le premier semestre de cette même année.

Simplication/décoration de textes

Simplification syntaxique

- *C'est aussi depuis le XVIIIe siècle le terme en usage pour désigner un clerc séculier ayant au moins reçu la tonsure.*
- *C'est aussi depuis le XVIIIe siècle le terme en usage.*

Simplication/décoration de textes

Simplification syntaxique

- *Ils ne sont pas caractérisés par leur profession comme dans la Bible : l'un pasteur, l'autre agriculteur.*
- *Ils ne sont pas caractérisés : l'un pasteur, l'autre agriculteur.*

Simplication/décoration de textes

Simplification lexicale

Objectifs :

- Rendre le texte plus facilement compréhensible au niveau lexical
- La complexité lexicale a plus d'impact sur la lisibilité et la compréhension [Arya *et al.*, 2011]

Simplication/décoration de textes

Simplification lexicale

Ressources :

- Données de la Toile pour exemplifier les entités nommées (personnes et lieux) [Lal & Ruger, 2002]
- Explication du vocabulaire pour les apprenants d'anglais [Burstein *et al.*, 2013]
- Définitions [Topac & Stoicu-Tivadar, 2013]
- Noms génériques
  [Thomas & Anderson, 2012, Abualhaija *et al.*, 2017]
- Vocabulaires spécifiquement acquis
  [Zhu *et al.*, 2010, Wubben *et al.*, 2012, Biran *et al.*, 2011, Glavas & Stajner, 2015, Kim *et al.*, 2016]

Simplication/décoration de textes

Simplification lexicale

- Compétition *SemEval 2012* [Specia *et al.*, 2012]
- Pour un texte court et un mot cible, et plusieurs substitutions possibles pour ce mot et satisfaisant le contexte, l'objectif était de trier ces substitutions selon leur degré de simplicité
- *Hitler committed terrible atrocities during the second World War.*
- candidats/synonyms : abomination, cruelty, enormity, violation
- bon choix : cruelty

Simplication/décoration de textes

Simplification lexicale

- Compétition *SemEval 2012* [Specia *et al.*, 2012]
- Pour un texte court et un mot cible, et plusieurs substitutions possibles pour ce mot et satisfaisant le contexte, l'objectif était de trier ces substitutions selon leur degré de simplicité
- *Hitler committed terrible atrocities during the second World War.*
- candidats/synonyms : abomination, cruelty, enormity, violation
- bon choix : cruelty
- *Hitler committed terrible cruelties during the second World War.*

Simplication/décoration de textes

Simplification lexicale

- Plusieurs stratégies :
  - lexique d'un corpus oral et de Wikipédia, n-grammes de Google, WordNet [Sinha, 2012] ;
  - longueur de mots, nombre des syllabes, information mutuelle et fréquence de mots [Jauhar & Specia, 2012] ;
  - fréquence dans Wikipédia, longueur de mots, n-grammes, complexité syntaxique des documents [Johannsen *et al.*, 2012] ;
  - n-grammes, fréquence dans Wikipédia, n-grammes de Google [Ligozat *et al.*, 2012] ;
  - WordNet et fréquences de mots [Amoia & Romanelli, 2012].

Simplication/décoration de textes

Simplification lexicale

- Remplacement, substitution

  {*lombalgie*, *douleurs lombaires*}
  {*hépatite*, *inflammation du foie*}

  1. *Les lombalgies inflammatoires provoquent une douleur de type inflammatoire*
     Les douleurs lombaires inflammatoires provoquent une douleur de type inflammatoire
  2. *Les lombalgies signifient les douleurs lombaires*
     Les douleurs lombaires signifient les douleurs lombaires
  3. *La lombalgie est une affection coûteuse pour le système de soins de santé et est un motif fréquent d'absentéisme.*
     La douleurs lombaires est une affection coûteuse pour le système de soins de santé et est un motif fréquent d'absentéisme.
  4. *Hépatite C : lutter contre le virus et ses résistances*
     Inflammation du foie C : lutter contre le virus et ses résistances

Simplication/décoration de textes

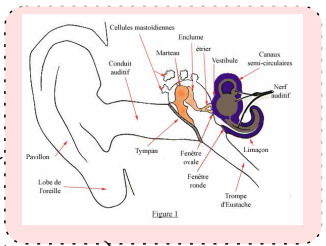
- Ajout d'informations
- Décoration

réparation

La myoplastie est une réfection chirurgicale d'un muscle.

inflammation de l'oreille

La pétrosite est une ostéite de la partie profonde du rocher (pyramide pétreuse)



presque toujours consécutive à une otite moyenne.

inflammation de l'oreille

affection de la peau (dermatose)

augmentation de volume

cellule de la peau

La mastocytose est une hyperplasie des mastocytes dont les manifestations

peuvent prédominer au niveau de la peau (Urticaire pigmentaire).

affection de la peau (dermatose)

Simplication/décoration de textes

Critères d'évaluation

- Simplicité
- Sémantique
- Grammaticalité

Simplication/décoration de textes

Bilan

- Domaine récent et peu travaillé
- Peu d'études dans la domaine médical
  - lexique spécifique
  - substitutions plus complexes
- Simplification lexicale et syntaxique séparément
- Différents types de destinataires

\+ Orthophonistes, TAL

? Médecins

Conclusion et Travaux futurs

Conclusion générale

- Différents aspects menant vers la simplification de textes
  - documents de spécialité
  - médecine
- Méthodes
  - diagnostic de textes
  - diagnostic de passages/mots non compréhensibles
- Ressources
  - plusieurs méthodes
  - évaluation des extractions
  - alignement avec les terminologies
- Simplification

Travaux futurs

- Améliorer la détection de mots incompréhensibles :
  - entités syntaxiquement complexes
  - formes fléchies et dérivées
  - orthographe
- Augmenter la couverture des paraphrases :
  - d'autres corpus
  - ressources supplétives plus couvrantes
  - d'autres méthodes pour extraire des paraphrases
  - gérer les paraphrases concurrentes
  - combinaison avec les images
- D'autres langues
- Simplification lexicale de textes médicaux
- Évaluation avec des utilisateurs


Abdaoui, A., Azé, J., Bringay, S., Grabar, N. & Poncelet, P. (2014).
Predicting medical roles in online health fora.
In *SLSP*, pp. 247–258.


Abualhaija, S., Miller, T., Eckle-Kohler, J., Gurevych, I. & Zimmermann, K.-H. (2017).
Metaheuristic approaches to lexical substitution and simplification.
In *EACL 2017*, pp. 1–11.


AMA (1999).
Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association.
JAMA, **281**(6), 552–7.


Amoia, M. & Romanelli, M. (2012).
SB : mmSystem - using decompositional semantics for lexical simplification.
In **SEM 2012*, pp. 482–486, Montréal, Canada.


Antoine, E. & Grabar, N. (2016).
Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert.
In *TALN 2016*.


Arya, D. J., Hiebert, E. H. & Pearson, P. D. (2011).
The effects of syntactic and lexical complexity on the comprehension of elementary science texts.
International Electronic Journal of Elementary Education, **4**(1), 107–125.


Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M. & Rynck, F. (2008).
Constitution de ressources pédagogiques numériques : le lexique des affects, In M. Loiseau, M. Abouzaïd, L. Buson, C. Cavalla, A. Djaroun, C. Dugua, A. Ghimenton, V. Goossens, T. Lebarbé, A. Nardy, F. Rinck & C. Surcouf, Eds., *Autour des langues et du langage : perspective pluridisciplinaire*, pp. 407–414.
Presses Universitaires de Grenoble.


Beigman Klebanov, B., Knight, K. & Marcu, D. (2004).
Text simplification for information-seeking applications.
In R. Meersman & Z. Tari, Eds., *On the Move to Meaningful Internet Systems 2004 : CoopIS, DOA, and ODBASE*. Berlin, Heidelberg : Springer, LNCS vol 3290.


Berland, G., Elliott, M., Morales, L., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Munoz, J., Puyol, J. & et al, M. L. (2001).
Health information on the internet. accessibility, quality, and readability in english ans spanish.
JAMA, **285**(20), 2612–2621.


Bhagat, R. & Hovy, E. (2013).
What is a paraphrase ?
Computational Linguistics, **39**(3), 463–472.


Biran, O., Brody, S. & Elhadad, N. (2011).
Putting it simply : a context-aware approach to lexical simplification.
In *ACL*.


Björnsson, H. & Härd af Segerstad, B. (1979).
Lix pä franska och tio andra språk.
Stockholm : Pedagogiskt centrum, Stockholms skolförvaltning.


Blake, C., Kampov, J., Orphanides, A., West, D. & Lown, C. (2007).
Query expansion, lexical simplification, and sentence selection strategies for multi-document summarization.
In *DUC*.


Brin-Henry, F. (2014).
Éducation thérapeutique du patient aphasique et son conjoint.
Rééducation orthophonique, **256**, 9–20.


Brouwers, L., Bernhard, D., Ligozat, A.-L. & Francois, T. (2014).
Syntactic sentence simplification for french.
In *PITR workshop*, pp. 47–56.


Burstein, J., Sabatini, J., Shore, J., Moulder, B. & Lentini, J. (2013).
A user study : Technology to increase teachers' linguistic awareness to improve instructional language support for English language learners.
In *Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pp. 20–29.


Candido, Jr., A., Maziero, E., Gasperin, C., Pardo, T. A. S., Specia, L. & Aluisio, S. M. (2009).
Supporting the adaptation of texts for poor literacy readers : a text simplification editor for Brazilian Portuguese.

In *EdAppsNLP '09 Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 34–42.

Cartoni, B. & Deléger, L. (2011).
Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes.
In *TALN*.

Chandrasekar, R. & Srinivas, B. (1997).
Automatic induction of rules for text simplification.
Knowledge Based Systems, **10**(3), 183–190.

Chen, P., Rochford, J., Kennedy, D. N., Djamasbi, S., Fay, P. & Scott, W. (2016).
Automatic text simplification for people with intellectual disabilities.
In *AIST*, pp. 1–9.

Chmielik, J. & Grabar, N. (2011).
Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques.
TAL, **51**(2), 151–179.

Claveau, V. & Kijak, E. (2014).
Generating and using probabilistic morphological resources for the biomedical domain.
In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3348–3354.

Cohen, J. (1960).

A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, **20**(1), 37–46.

Collaboration, C. (2009).
Cochrane : systematic review of biomedical literature.
Cochrane Collaboration.
www.cochrane.org.

Côté, R. A., Rothwell, D. J., Palotay, J. L., Beckett, R. S. & Brochu, L. (1993).
The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International.
Northfield : College of American Pathologists.

De Belder, J. & Moens, M.-F. (2010).
Text simplification for children.
In *Workshop on Accessible Search Systems of SIGIR*, pp. 1–8.

Deléger, L. & Zweigenbaum, P. (2008).
Paraphrase acquisition from comparable medical corpora of specialized and lay texts.
In *AMIA 2008*, pp. 146–50.

d'Ivernois, J.-F., Gagnayre, R. & *et al* (2011).
Compétences d'adaptation à la maladie du patient : une proposition [*The patient's psychosocial skills : a proposal*].
Educ Ther Patient/Ther Patient Educ, **3**(2), S201–S205.

Elhadad, N. & Sutaria, K. (2007).

Mining a lexicon of technical terms and lay equivalents.
In *BioNLP*, pp. 49–56.

Flesch, R. (1948).
A new readability yardstick.
Journ Appl Psychol, **23**, 221–233.

Glasgow, R. E., Kurz, D., King, D., Dickman, J. M., Faber, A. J., Halterman, E., Woolley, T., Toobert, D. J. & et al, L. A. S. (2012).
Twelve-month outcomes of an internet-based diabetes self-management support program.
Patient Education and Communication, **87**(1), 81–92.

Glavas, G. & Stajner, S. (2015).
Simplifying lexical simplification : Do we need simplified corpora ?
In *ACL-COLING*, pp. 63–68.

Goeuriot, L., Grabar, N. & Daille, B. (2007).
Caractérisation des discours scientifique et vulgarisé en français, japonais et russe.
In *TALN*, pp. 93–102.

Golay, A., Lagger, G. & Giordan, A. (2007).
Motivating patient with chronic diseases.
Journ of Med and the Person, **5**(2), 57–63.

Grabar, N., Chauveau Thoumelin, P. & Dumonet, L. (2015).
Study of subjectivity in the medical discourse : Uncertainty and emotions.
Advances in Knowledge Discovery and Management, **5**, 33–54.


Grabar, N. & Hamon, T. (2016).
Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux.
TAL, **57**(1), 85–109.


Grabar, N. & Krivine, S. (2007).
Application of cross-language criteria for the automatic distinction of expert and non expert online health documents.
In *Springer LNAI (AIME 2007)*, pp. 252–256.


Grabar, N., Krivine, S. & Jaulent, M. (2007).
Classification of health webpages as expert and non expert with a reduced set of cross-language features.
In *AMIA*, pp. 284–288.


Gross, O. & Gagnayre, R. (2013).
Hypothèse d'un modèle théorique du patient-expert et de l'expertise du patient : processus d'élaboration.
Recherches qualitatives, **15**(HS), 147–165.


Guilbert, M. (2014).
C'est grave docteur ?
Europe : Les Éditions de l'Opportun.


Gunning, R. (1973).
The art of clear writing.
New York, NY : McGraw Hill.


Jauhar, S. & Specia, L. (2012).

UOW-SHEF : SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features.
In **SEM 2012*, pp. 477–481, Montréal, Canada.

Johannsen, A., Martínez, H., Klerke, S. & Søgaard, A. (2012).
Emnlp@cph : Is frequency all there is to simplicity ?
In **SEM 2012*, pp. 408–412, Montréal, Canada.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C. & Gonzalez, G. (2009).
Towards effective sentence simplification for automatic processing of biomedical text.
In *NAACL HLT 2009*, pp. 177–180.

Kim, Y.-S., Hullman, J., Burgess, M. & Adar, E. (2016).
Simplescience : Lexical simplification of scientific terminology.
In *EMNLP*, pp. 1–6.

Kincaid, J., Fishburne, R. J., Rogers, R. & Chissom, B. (1975).
Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.
Technical report, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kokkinakis, D. & Toporowska Gronostaj, M. (2006).
Comparing lay and professional language in cardiovascular disorders corpora.
In A. Pham T., James Cook University, Ed., *WSEAS Transactions on BIOLOGY and BIOMEDICINE*, pp. 429–437.


Lal, P. & Ruger, S. (2002).
Extract-based summarization with simplification.
In *ACL*.


Laurent, D., Nègre, S. & Séguéla, P. (2009).
L'analyseur syntaxique Cordial dans Passage.
In *TALN 2009*.


Le Bot, M.-C., Schuwer, M. & Élisabeth Richard (dir.) (2008).
La reformulation : Marqueurs linguistiques – Stratégies énonciatives.
Rennes : Rivages linguistiques.


Leroy, G., Kauchak, D. & Mouradi, O. (2013).
A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty.
Int J Med Inform, **82**(8), 717–730.


Levenshtein, V. I. (1966).
Binary codes capable of correcting deletions, insertions and reversals.
Soviet physics. Doklady, **707**(10).


Ligozat, A., Grouin, C., Garcia-Fernandez, A. & Bernhard, D. (2012).
Annlor : A naïve notation-system for lexical outputs ranking.
In **SEM 2012*, pp. 487–492.


Lindberg, D., Humphreys, B. & McCray, A. (1993).
The unified medical language system.
Methods Inf Med, **32**(4), 281–291.


Max, A. (2008).
Local rephrasing suggestions for supporting the work of writers.
In *GOTAL*, pp. 324–335.


McCray, A. (2005).
Promoting health literacy.
J of Am Med Infor Ass, **12**, 152–163.


Miller, T., Leroy, G., Chatterjee, S., Fan, J. & Thoms, B. (2007).
A classifier to evaluate language specificity of medical documents.
In *HICSS*, pp. 134–140.


Namer, F. (2000).
FLEMM : un analyseur flexionnel du français à base de règles.
Traitement automatique des langues (TAL), **41**(2), 523–547.


Namer, F. (2003).
Automatiser l'analyse morpho-sémantique non affixale : le système DériF.
Cahiers de Grammaire, **28**, 31–48.


Paetzold, G. H. & Specia, L. (2016).
Benchmarking lexical simplification systems.
In *LREC*, pp. 3074–3080.


Patel, V., Branch, T. & Arocha, J. (2002).
Errors in interpreting quantities as procedures : The case of pharmaceutical labels.
Int journ med inform, **65**(3), 193–211.


Péry-Woodley, M. & Rebeyrolle, J. (1998).
Domain and genre in sublanguage text : definitional microtexts in three corpora.
In *LREC*, pp. 987–992.


Poprat, M., Markó, K. & Hahn, U. (2006).
A language classifier that automatically divides medical documents for experts and health care consumers.
In *MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics*, pp. 503–508, Maastricht.


Risk, A. & Dzenowagis, J. (2001).
Review of internet information quality initiatives.
Journal of Medical Internet Research, **3**(4).


Schmid, H. (1994).
Probabilistic part-of-speech tagging using decision trees.
In *ICNMLP*, pp. 44–49, Manchester, UK.


Schwartz, A. S. & Hearst, M. A. (2003).
A simple algorithm for identifying abbreviation definitions in biomedical text.
In *Pacific Symposium on Biocomputing*, pp. 451–456.


Siddharthan, A. (2006).
Syntactic simplification and text cohesion.
Research on Language & Computation, **4**(1), 77–109.


Sinha, R. (2012).
Unt-simprank : Systems for lexical simplification ranking.
In **SEM 2012*, pp. 493–496.


Son, J. Y., Smith, L. B. & Goldstone, R. L. (1008).
Simplicity and generalization : Short-cutting abstraction in children's object categorizations.
Cognition, **108**, 626–638.


Specia, L., Jauhar, S. & Mihalcea, R. (2012).
Semeval-2012 task 1 : English lexical simplification.
In **SEM 2012*, pp. 347–355.


Stymne, S., Tiedemann, J., Hardmeier, C. & Nivre, J. (2013).
Statistical machine translation with readability constraints.
In *NODALIDA*, pp. 1–12.


Sørensen, M. H. (1996).
Turchin's Supercompiler Revisited - An operational theory of positive information propagation.
Master thesis, University of Copenhagen, Copenhagen, Denmark.


Thomas, S. R. & Anderson, S. (2012).
Wordnet-based lexical simplification of a document.
In *KONVENS*, pp. 80–88.


Topac, V. & Stoicu-Tivadar, V. (2013).
Patient empowerment by increasing the understanding of medical language for lay users.
Methods Inf Med, **52**(5), 454–62.


Tran, T., Chekroud, H., Thiery, P. & Julienne, A. (2009).
Internet et soins : un tiers invisible dans la relation médecine/patient ?

Ethica Clinica, **53**, 34–43.


Vickrey, D. & Koller, D. (2008).
Sentence simplification for semantic role labeling.
In *ACL-HLT*, pp. 344–352.


Štajner, S. & Popović, M. (2016).
Can text simplification help machine translation ?
Baltic J. Modern Computing, **4**(2), 230–242.


Vu, T. T., Tran, G. B. & Pham, S. B. (2014).
Learning to simplify children stories with limited data.
In L. . Springer, Ed., *Intelligent Information and Database Systems*, pp. 31–41.


Wang, Y. (2006).
Automatic recognition of text difficulty from consumers health information.
In IEEE, Ed., *Computer-Based Medical Systems*, pp. 131–136.


Wei, C.-H., Leaman, R. & Lu, Z. (2014).
Simconcept : A hybrid approach for simplifying composite named entities in biomedicine.
In *BCB '14*, pp. 138–146.


Williams, M., Parker, R., Baker, D., Parikh, N., Pitkin, K., Coates, W. & Nurss, J. (1995).
Inadequate functional health literacy among patients at two public hospitals.
JAMA, **274**(21), 1677–1682.


Witten, I. & Frank, E. (2005).

Data mining : Practical machine learning tools and techniques.
Morgan Kaufmann, San Francisco.

Wubben, S., van den Bosch, A. & Krahmer, E. (2012).
Sentence simplification by monolingual machine translation.
In *ACL*, pp. 1015–1024.

Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J. & Browne, A. C. (2006).
Exploring lexical forms : first-generation consumer health vocabularies.
In *AMIA 2006*, pp. 1155–1155.

Zeng-Treiler, Q., Kim, H., Goryachev, S., Keselman, A., Slaugther, L. & Smith, C. (2007).
Text characteristics of clinical reports and their implications for the readability of personal health records.
In *MEDINFO*, pp. 1117–1121, Brisbane, Australia.

Zheng, W., Milios, E. & Watters, C. (2002).
Filtering for medical news items using a machine learning approach.
In *AMIA*, pp. 949–53.

Zhu, Z., Bernhard, D. & Gurevych, I. (2010).
A Monolingual Tree-based Translation Model for Sentence Simplification.
In *COLING 2010*, pp. 1353–1361.