Extraction automatique de paraphrases grand public pour les termes médicaux

Natalia Grabar[1], Thierry Hamon[2]
Présenté par Iris Eshkol-Taravella

(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr
(2) LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr

TALN 2015, Caen

Contexte

- Disponibilité des informations de santé
  - recherche scientifique, médias sociaux, documents cliniques, télé, radio, nouvelles
- Termes très spécifiques
  - *myocarde, cholecystectomie, blépharospasme, alexitymie, apendicectomie, desmorrhexie, lombalgie*
- Compréhension souvent difficile (AMA, 1999; Mccray, 2005; Eysenbach, 2007):
  - préparation et prise de médicaments (Patel et al., 2002);
  - notices de médicaments, brochures et consensus informés (Williams et al., 1995);
  - informations de santé en ligne (Berland et al., 2001; Hargrave et al., 2003; Kusec et al., 2004)
- Effet négatif sur le processus de soins médicaux (Tran et al., 2009)

Objectifs

- Acquisition automatique de paraphrases pour les termes médicaux
- Composition néoclassique
  - *myocarde, cholecystectomie, blépharospasme, alexitymie, apendicectomie, desmorrhexie, lombalgie*
- Difficulté: bases supplétives: {*myo*, *muscle*}, {*cardia*, *cœur*}
- Paraphrases:
  - {*myocardiaque*, *muscle du cœur*}
  - {*cholecystectomie*, *ablation de la vésicule biliaire*}
- Corpus de la langue générale

Travaux existants

- Lisibilité: facilité de compréhension d'un texte
  - mesures classiques (Flesch, 1948; Gunning, 1973; Dubay, 2004)
  - mesures computationnelles (Wang et al., 2006; Zeng et al., 2007; Leroy et al., 2008)
- Simplification lexicale: rendre un texte plus facile à comprendre (Specia et al., 2012; Siddharthan, 2015)
  - *The police enquiry also discovered evidence that he has successfully intercepted voicemail messages belong to Rebekah Brooks.*
  - *The enquiry found proof that he has intercepted messages belong to Rebekah Brooks.*
- Ressources pour la simplification
  - {*myocardial infarction*, *heart attack*} (Zeng et al., 2006)
  - {*apports caloriques*, *apport en calories*} (Deleger et al., 2008)
- Décomposition de composés néoclassiques (McCray et al., 1988; Namer, 2003; Loginova et al., 2013; Claveau et al., 2014)

Données linguistiques

Termes médicaux

- Snomed International (Cote, 1997), partie française d'UMLS (Lindberg et al., 1993)
  - mots des termes
  - pas de nombres

Données linguistiques

Corpus

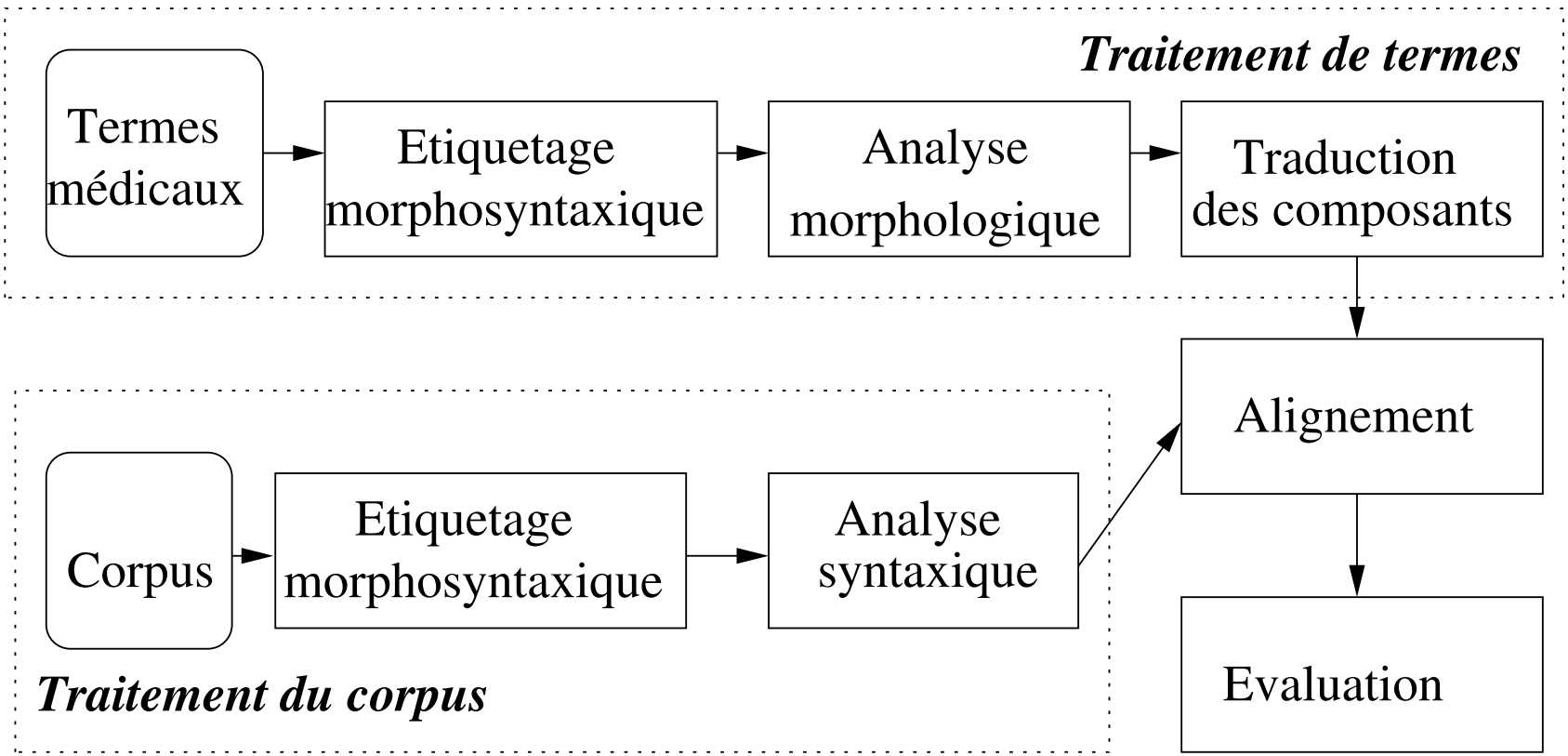
- Wikipédia, Portail de la Médecine
  - version de janvier 2015
  - 18 434 articles
  - 15 235 219 occurrences

Données linguistiques

Ressources linguistiques

- Ressources morphologiques (155 468)
  - {*aorte*, *aortique*}, {*aortique*, *aortiques*}
- Ressources de synonymes (4 914)
  - {*embolie*, *thrombose*}, {*tumeur*, *fibrome*}
- Ressources supplétives (1 022)
  - {*base supplétive*, *mot du français*}
  - {*andr*, *mâle*}, {*ectomie*, *ablation*}, {*myo*, *muscle*}

Méthode


Traitement de termes
Termes médicaux
Etiquetage morphosyntaxique
Analyse morphologique
Traduction des composants
Alignement
Corpus
Etiquetage morphosyntaxique
Analyse syntaxique
Traitement du corpus
Evaluation

Méthode

1. Traitement de termes médicaux

- Étiquetage morpho-syntaxique et lemmatisation `Cordial` (Laurent et al., 2009)
  - *myocardique/A*, *cholécystectomie/N*
- Analyse morphologique `DériF` (Namer, 2009)
  - *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*

Méthode

2. Traitement du corpus

- `Cordial` (Laurent et al., 2009):
  - étiquetage morpho-syntaxique et lemmatisation
  - analyse syntaxique
- Définir les frontières des syntagmes

Méthode

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

Méthode

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
- Baseline: contextes définitoires

Résultats

Extraction

- 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)

Résultats

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

Résultats

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

Discussion

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*

Discussion

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)

Discussion

Extraction de paraphrases et leur évaluation

- Relations sémantique 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

Discussion

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)

Discussion

Comparaison avec les contextes définitoires

- Extraction:
  - 4 patrons (Pery et al., 1998)
    - *désigne, est un, est appelé, peut être défini comme*
  - 2 037 contextes définitoires
  - 1 286 termes uniques
  - termes composés, affixés, morphologiquement simples
- Évaluation:
  - précision stricte: 52,5 %
  - précision lâche: 68 %
- 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.*
- Résultats comparables (et complémentaires)

Discussion

Comparaison avec les travaux existants

	type terme	nb. para	précision
(Elhadad et al., 2007)	tous	152	58
(Deleger et al., 2008)	m-synt.	65, 82	67, 60
(Cartoni et al., 2011)	m-synt.	109	66
notre travail	composés	1 128	76, 86

- morpho-syntaxique:
  - {*consommation régulière*, *consommer de façon régulière*}
- performances comparables, meilleure couverture
- `DériF` (Namer, 2009):
  - 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*

Discussion

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 et al., 2014)

Conclusion

- Paraphrases grand public pour les termes médicaux
- Composés néoclassiques
- Corpus non spécialisé: Wikipédia
- Résultats:
  - jusqu'à 1 128 termes
  - 2 089 avec les définitions
- Précision moyenne:
  - toutes les expériences: 76 %
  - sans synonymes: 86 %

Travaux futurs

- Augmenter la couverture:
  - d'autres corpus
  - ressources supplétives plus couvrantes
  - d'autres méthodes pour extraire des paraphrases
- Termes complexes syntaxiquement:
  - *vaporisateur hypodermique, fistule trachéo-œsophagienne, cardiopathie artérioscléreuse*
- D'autres langues
- Simplification lexicale de textes médicaux