

Analyse de corpus pour la simplification automatique de textes médicaux

Rémi Cardon

CNRS, UdLille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
remi.cardon@univ-lille.fr

M2 LITL, Université de Toulouse Jean Jaurès



Plan

- 1 La simplification automatique de textes
- 2 Présentation du projet CLEAR
- 3 Le corpus CLEAR
- 4 Substitution lexicale
- 5 Alignement de phrases
- 6 Conclusion

Simplification automatique de textes – Objectifs

- Rendre l'information accessible aux humains
 - Enfants ou adultes ayant des difficultés de lecture (Bosch *et al.*, 2008; De Belder *et al.*, 2010; Vu *et al.*, 2014; Paetzold & Specia, 2016a)
 - Étrangers (Paetzold & Specia, 2016b)
 - Adultes ayant des troubles neurocognitifs (Chen *et al.*, 2016)
 - Grand public (textes spécialisés) (Arya *et al.*, 2011)
 - ...
- Pré-traitement de données pour le TAL
 - Analyse syntaxique (Chandrasekar & Srinivas, 1997)
 - Résumé automatique (Blake *et al.*, 2007)
 - Traduction automatique (Stymne *et al.*, 2013; Štajner & Popović, 2016)
 - Extraction / Recherche d'information (Beigman Klebanov *et al.*, 2004)
 - ...

Simplification automatique de textes – Méthodes

- Modifier un texte pour le rendre accessible, tout en préservant son sens
- Principalement deux versants :
 - Simplification lexicale : remplacer les termes ou expressions complexes par des équivalents accessibles
 - Simplification syntaxique : retravailler les structures de phrases
- Schéma d'annotation pour le processus de simplification (Brunato *et al.*, 2014)
 - Découpage, fusion, réorganisation de phrases
 - Insertion / Effacement (verbes, sujets, autres)
 - Transformation (substitution lexicale, remplacement des anaphores, modification des traits verbaux...)
- Pas de consensus pour l'évaluation d'un système de simplification
- Nécessite des corpus parallèles avec différents degrés de complexité

Travail existant

- Corpus parallèles :
 - English, Spanish, Italian, Brazilian Portuguese, Danish
(Chandrasekar & Srinivas, 1997; Bott *et al.*, 2014; Brunato *et al.*, 2014; Caseli *et al.*, 2009; Klerke & Søggaard, 2012)
 - Pas toujours accessibles librement
- Corpus comparables :
 - Anglais : Simple English Wikipedia - English Wikipedia (SEW-EW), Newsela
 - Autres : Révisions dans Simple English Wikipedia, articles scientifiques ou romans simplifiés
- Pas de ressource en français

Le projet CLEAR

- *Communication, Literacy, Education, Accessibility, Readability*
- Projet ANR (Agence Nationale pour la Recherche)
ANR-17-CE19-0016-01, 2018-2020
- Objectif : Proposer des méthodes pour la création de ressources et d'un prototype pour la simplification de textes médicaux en français.
- Porté par STL (Natalia Grabar)
- Partenaires :
 - STL (Savoirs, Textes, Langage)
 - LIMSI (Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur)
 - LEPS (Laboratoire de Pédagogie de la Santé)
 - AFH (Association Française des Hémophiles)
 - Synapse Développement

Le corpus CLEAR

- Corpus comparable médical en français, pour la simplification de textes biomédicaux
- Trois sous-corpus :
 - Articles encyclopédiques
 - Notices de médicaments
 - Articles scientifiques
- Paires de documents
 - Même sujet
 - Deux niveaux de complexité

Articles Encyclopédiques

- Publiés par la Fondation Wikimedia
- Wikipedia: s'adresse au grand public

(2 186 891 tokens, 19 287 lemmes)

- Vikidia: s'adresse aux enfants de 8 à 13 ans

(183 051 tokens, 3 117 lemmes)

- Rédigés indépendamment
- Articles du portail de la médecine
- 2*575 articles
- Exemples :

- La luette ou uvule est un appendice conique situé au fond de la **cavité buccale**. La luette est un organe de 10 à 15 millimètres de long. Elle est constituée d'un tissu membraneux et musculaire.
- La luette ou uvule est un appendice conique situé au fond de la **bouche**. C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15mm de long, qui pend à la partie moyenne du voile du palais.

Notices de médicament

- Publiés par le Ministère de la Santé
- Résumés des caractéristiques produit (RCP), pour les praticiens

(51 705 111 tokens, 43 515 lemmes)

- Notices, pour les patients

(33 116 119 tokens, 25 725 lemmes)

- 2*11 800 notices

- Exemples :

- **hypersensibilité** à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des **difficultés de déglutition** en raison du risque d'inhalation bronchique et de pneumopathie lipoïde.
- si vous avez une **allergie** à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des **difficultés pour avaler** en raison du risque d'inhalation de la paraffine liquide qui entraîne une pneumopathie lipoïde.

Articles scientifiques

- Publiés par la Fondation Cochrane
- Revues de littérature médicale pour les praticiens

(2 804 335 tokens, 11 558 lemmes)

- Versions simplifiées manuellement par la fondation pour le grand public

(1 491 243 tokens, 7 567 lemmes)

- 2*3 815 résumés
- Exemples :
 - L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant.
 - L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant.

Résumé

	<i>docs</i>	<i>tokens_{tech}</i>	<i>tokens_{acc}</i>	<i>total tokens</i>	<i>lemmes_{tech}</i>	<i>lemmes_{acc}</i>
<i>Wiki</i>	575*2	2,293,078	197,672	2,490,750	19 287	3 117
<i>Médicaments</i>	11,800*2	52,313,126	33,682,889	85,996,015	43 515	25 725
<i>Cochrane</i>	3,815*2	2,840,003	1,515,051	4,355,054	11 558	7 567
<i>Total</i>	16,190	57,446,207	35,395,612	92,841,819		

- 16 190 paires de documents
- Technique : 57M tokens
- Simple : 35M tokens
- Total : 92M tokens
- Plus grande diversité lexicale côté technique

Substitution lexicale

- Remplacer des termes techniques par des expressions équivalentes plus simples
- Plusieurs étapes (Shardlow, 2014) :
 - Repérage des termes difficiles
 - Mesures de lisibilité
 - Génération des candidats
 - Choix parmi les candidats
 - Substitution

Substitution lexicale – Ressources

- Lexique Snomed International (Côté, 1996)
 - 151 104 termes médicaux
 - 11 types sémantiques (maladies/anomalies, actes médicaux, produits chimiques, anatomie...)
 - Utilisé pour estimer la couverture du domaine médical d'un texte / corpus
- GLAWI (Sajous & Hathout, 2015)
 - Version XML du Wiktionary
 - Récupération des termes des catégories suivantes : *anat, chirurgie, génétique, maladie(s), médecine, médicaments, microbiologie, neurologie, pathologie, pédologie, pharmacologie, physiologie, squelette, virologie*
 - 8 012 entrées (médecine : 4 967, anat : 1 925)
 - Synonymes et hyperonymes : 25% du lexique constitué (2,42/terme en moyenne)
 - Essentiellement des termes simples (vs. unités polylexicales)
 - Fournit le paradigme flexionnel (lemmes + formes fléchies)

Analyse lexicale

- Nombre de termes uniques de la Snomed International

<i>Corpus</i>	<i>Termes</i>
Encyclopédie simplifié	1 635
Encyclopédie technique	3 999
Médicaments simplifié	2 700
Médicaments technique	3 332
Cochrane simplifié	2 316
Cochrane technique	2 505

Analyse lexicale

- Ratio des occurrences de termes de la Snomed International

<i>Ratio</i>	<i>Valeur</i>
Cochrane : simplifié / technique	0.60
Cochrane : technique / simplifié	1.66
Médicaments : simplifié / technique	0.62
Médicaments : technique / simplifié	1.62
Encyclopédie : simplifié / technique	0.10
Encyclopédie : technique / simplifié	9.67

Substitution lexicale – Méthode

- Test sur phrases des corpus techniques
- GLAWI fournit les candidats à la substitution
- Filtrage de GLAWI sur Vikidia : si entrée ou synonyme / hyperonyme est dans Vikidia, terme considéré comme simple
- Si une phrase technique contient une entrée avec synonyme / hyperonyme et que cette entrée n'est pas dans Vikidia, on utilise le synonyme / hyperonyme pour substitution

Substitution lexicale – Évaluation

- Évaluation manuelle
- Critères (Biran *et al.*, 2011; Wubben *et al.*, 2012) :
 - Grammaticalité : la phrase est-elle encore grammaticalement valide ?
 - Sémantique : la phrase a-t-elle gardé le même sens ?
 - Simplicité : la phrase est-elle plus simple à comprendre ?

Substitution lexicale – Évaluation

- Application de la méthode sur 10 documents de chaque corpus (7 892 phrases)
- 26 phrases modifiées
 - *prescription – traitement*
 - *transformation – traitement*
 - Encyclopédie : 5 057
 - Médicaments : 2 456
 - Cochrane : 379

<i>Critère</i>	<i>% Méthode</i>	<i>Devlin</i>	<i>Biran</i>
Grammaticalité	70%	70.23%	77.91%
Simplicité	14.46%	46.43%	75.58%
Sémantique	18.51%	55.95%	46.43%

Exemples – Sans la contrainte d'absence dans Vikidia

Avant substitution

La nourriture n'a pas d'effet sur l'absorption d'anastrozole.

Vous devez discuter avec votre médecin sur les risques et les options de traitement.

Votre médecin peut vous prescrire un médicament visant à prévenir ou traiter cette perte osseuse.

Après substitution

La nourriture n'a pas d'effet sur l'ingestion d'anastrozole.

Vous devez discuter avec votre médecin sur les risques et les options de prescription.

Votre médecin peut vous prescrire un médicament visant à prévenir ou soigner cette perte osseuse.

Avant substitution

Un abcès est une accumulation de pus sous la peau ou à l'intérieur du corps.

Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la main).

Après substitution

Un abcès est une accumulation de pus sous la peau ou à l'intérieur du mort.

Syndrome du canal carpien (fourmillement, douleur, sensation de froid, faiblesse dans certaines parties de la pince).

Alignement automatique

- Objectif : accroître le corpus de phrases parallèles
- Alignement manuel : données d'entraînement
- Unité de traitement : paire de phrases
- Trop peu de données pour les réseaux de neurones
- Approche : tâche de classification binaire (paire alignée / non-alignée)
- Données :
 - paires alignées : résultat de l'alignement manuel
 - paires non alignées : appariement aléatoire

Données de référence

- Alignement manuel de phrases parallèles
- 39 paires de documents sélectionnées aléatoirement
 - 2*14 articles encyclopédiques
 - 2*12 notices de médicament
 - 2*13 résumés Cochrane
- Deux annotateurs
 - Annotation indépendante

Critères d'alignement

- Alignement :
 - Équivalence – sens identique ou presque identique :
 - Une gêne visuelle passagère peut être ressentie après **instillation** du collyre.
 - Une gêne visuelle passagère peut être ressentie après l'**administration** du collyre.
 - Inclusion – le sens d'une phrase se retrouve intégralement dans une autre :
 - **La maladie de Charcot est l'autre nom de la sclérose latérale amyotrophique.**
 - Il est le découvreur de **la sclérose latérale amyotrophique, ou maladie de Charcot**, une maladie neurodégénérative.

Critères d'alignement

- Non-alignement :
 - Phrases identiques, ou qui varient seulement par la ponctuation ou les mots grammaticaux :
 - Effets sur l'aptitude à conduire des véhicules **ou** à utiliser des machines
 - Effets sur l'aptitude à conduire des véhicules **et** à utiliser des machines
 - Intersection – deux phrases qui partagent du sens mais qui apportent chacune une information qui leur est propre :
 - Une faiblesse musculaire (hypotonie axiale), des difficultés d'alimentation (troubles de la succion entraînant une faible prise de poids), une hyperexcitabilité, une agitation ou des tremblements **peuvent survenir chez le nouveau-né**, ces troubles étant réversibles.
 - Un traitement en fin de grossesse par benzodiazépines même à faibles doses, peut être responsable **chez le nouveau-né de signes d'imprégnation tels qu'hypotonie axiale, troubles de la succion entraînant une faible prise de poids.**

Résultats de l'alignement

corpus	doc.	Technique					Simple					Taux d'alignement (%)	
		source		aligné			source		aligné			alignement (%)	
		ph.	tokens	ph.	tokens		ph.	tokens	ph.	tokens		tech.	simp.
Wiki	14*2	2,494	36,002	49	1,100		238	2,659	49	853		2	21
Médicaments	12*2	4,416	44,709	502	5,751		2,736	27,820	502	10,398		18	11
Cochrane	13*2	553	8,854	112	3,166		263	4,688	112	3,306		20	43

- All (kappa de Cohen) avant session de consensus : 0.76
- 663 paires alignées
- Différences substantielles entre les corpus
- Répartition :
 - 40% d'équivalence
 - 40% d'inclusion technique dans simple
 - 20% d'inclusion simple dans technique

Alignement automatique

- Descripteurs des paires de phrases :
 - BL :
 - Nombre de mots communs hors stopwords
 - Nombre de stopwords communs
 - Pourcentage de mots d'une phrase présents dans l'autre
 - Différence de longueur entre les deux phrases (en mots)
 - Différence de longueur moyenne des mots entre les deux phrases
 - N :
 - Nombre total de bigrammes et trigrammes communs (en caractères)
 - S :
 - Mesures de similarité : cosine, Dice, Jaccard
 - L :
 - Distance d'édition (Levenshtein) entre les deux phrases (caractères, et mots)

Alignement automatique

- Choix de l'algorithme (Scikit-learn ¹) :
 - 663 paires alignées, 800 non alignées
 - Train 2/3, Test 1/3

<i>Classifieur</i>	<i>R</i>	<i>P</i>	<i>F1</i>	<i>MSE</i>
<i>Perceptron</i>	0.87	0.84	0.84	0.63
<i>MLP</i>	0.87	0.87	0.86	0.53
<i>LDA</i>	0.90	0.90	0.90	0.40
<i>QDA</i>	0.89	0.89	0.89	0.45
LogReg	0.93	0.93	0.93	0.30
<i>SGD</i>	0.87	0.84	0.84	0.84
<i>LinSVM</i>	0.81	0.81	0.81	0.74

- → Régression logistique

Alignement automatique

- Évaluation des descripteurs :

<i>Descripteurs</i>	<i>R</i>	<i>P</i>	<i>F1</i>	<i>MSE</i>
<i>BL</i>	0.87	0.87	0.86	0.54
<i>S</i>	0.84	0.84	0.84	0.64
<i>L</i>	0.79	0.78	0.78	0.86
<i>N</i>	0.89	0.88	0.88	0.48
<i>L+S</i>	0.88	0.88	0.88	0.48
<i>L+N</i>	0.91	0.91	0.91	0.37
<i>S+N</i>	0.91	0.91	0.91	0.37
<i>BL+L</i>	0.90	0.90	0.90	0.40
<i>BL+S</i>	0.89	0.89	0.89	0.46
<i>BL+N</i>	0.91	0.91	0.91	0.35
<i>BL+L+S</i>	0.90	0.90	0.90	0.40
BL+L+N	0.93	0.93	0.93	0.29
<i>BL+S+N</i>	0.91	0.91	0.91	0.36
<i>L+S+N</i>	0.91	0.91	0.91	0.36
BL+L+S+N	0.93	0.93	0.93	0.29

Alignement automatique

- Application sur les données réelles
- 663 paires alignées, 540 000 paires non alignées
 - Score global : $P : 1.00$, $R : 1.00$, $F1 : 1.00$
 - Score sur paires alignées : $P : 0.62$, $R : 0.14$, $F1 : 0.22$
- Problème du déséquilibre des données
- Pistes :
 - Adapter l'algorithme
 - Adapter les données

Alignement automatique – STS Monolingue

- Application aux données STS (SemEval 2013, 2014, 2015)
- Entraînement et test sur données STS

<i>Données Score</i>	<i>R</i>	<i>P</i>	<i>F1</i>	<i>MSE</i>	<i>TP</i>
<i>STS2012 2.5</i>	0.82	0.82	0.82	0.71	477
<i>STS2012 3.5</i>	0.74	0.74	0.74	1.04	277
<i>STS2012 4.5</i>	0.79	0.81	0.78	0.74	37
<i>STS2013 2.5</i>	0.73	0.73	0.73	1.09	176
<i>STS2013 3.5</i>	0.78	0.78	0.78	0.87	96
<i>STS2013 4.5</i>	0.89	0.93	0.90	0.29	2
<i>STS2014 2.5</i>	0.75	0.76	0.75	0.97	653
<i>STS2014 3.5</i>	0.70	0.71	0.71	1.17	306
<i>STS2014 4.5</i>	0.89	0.93	0.90	0.29	2

Alignement automatique – STS Cross-lingue

- Application aux données STS (SemEval 2013, 2014, 2015)
- Entraînement sur données CLEAR et test sur données STS

<i>Données Score</i>	<i>R</i>	<i>P</i>	<i>F1</i>	<i>MSE</i>	<i>TP</i>
<i>STS2012 2.5</i>	0.83	0.81	0.82	0.19	1378
<i>STS2012 3.5</i>	0.74	0.72	0.71	0.28	1035
<i>STS2012 4.5</i>	0.81	0.49	0.52	0.51	413
<i>STS2013 2.5</i>	0.74	0.74	0.74	0.26	523
<i>STS2013 3.5</i>	0.78	0.73	0.74	0.27	396
<i>STS2013 4.5</i>	0.92	0.57	0.67	0.43	88
<i>STS2014 2.5</i>	0.74	0.72	0.73	0.28	1688
<i>STS2014 3.5</i>	0.72	0.69	0.69	0.31	1216
<i>STS2014 4.5</i>	0.88	0.54	0.61	0.46	384

Conclusion

- Difficultés liés au domaine :
 - Plus faible recouvrement au niveau des mots ou des caractères que dans la langue générale :
 - *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.*
 - *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.*
 - Besoin de ressources spécifiques au domaine

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

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.

BIRAN O., BRODY S. & ELHADAD N. (2011). Putting it simply: a context-aware approach to lexical simplification. In ACL, Ed., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, p. 496–501.

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*.

- BOSCH S., PRETORIUS L. & FLEISCH A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, **17**(2), 66–88.
- BOTT S., SAGGION H. & MILLE S. (2014). Text simplification tools for spanish. In *LREC 2014*, p. 1–7.
- BRUNATO D., DELL'ORLETTA F., VENTURI G. & MONTEMAGNI S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *CLICIT*, p. 87–92.
- CASELI H. M., PEREIRA T. F., SPECIA L., PARDO T. A. S., GASPERIN C. & ALUISIO S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *CICLING*, p. 1–12.
- 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*, p. 1–9.

- CÔTÉ R. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale*, v3.4. Université de Sherbrooke, Sherbrooke, Québec.
- DE BELDER J., DESCHACHT K. & MOENS M.-F. (2010). Lexical simplification. In *ITEC*. 1-4.
- KLERKE S. & SØGAARD A. (2012). DSim, a Danish parallel corpus for text simplification. In *LREC*, p. 4015–4018.
- PAETZOLD G. H. & SPECIA L. (2016a). Benchmarking lexical simplification systems. In *LREC*, p. 3074–3080.
- PAETZOLD G. H. & SPECIA L. (2016b). Unsupervised lexical simplification for non-native speakers. In *AAAI Conference on Artificial Intelligence*, p. 3761–3767.
- PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COUNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

- SAJOUS F. & HATHOUT N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the eLex 2015 conference*, p. 405–426, Herstmonceux, England.
- SHARDLOW M. (2014). A survey of automated text simplification. *Int J Advanced Computer Science and Applications*, **1**, 1–13.
- STYMNE S., TIEDEMANN J., HARDMEIER C. & NIVRE J. (2013). Statistical machine translation with readability constraints. In *NODALIDA*, p. 1–12.
- 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*, p. 31–41.
- WUBBEN S., VAN DEN BOSCH A. & KRAHMER E. (2012). Sentence simplification by monolingual machine translation. In *ACL*, p. 1015–1024.