Acquisition of Expert/Non-expert Vocabulary from Reformulations

Edwige ANTOINE[a,] and Natalia GRABAR[a1]
[a] *Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France*

Abstract. Technical medical terms are complicated to be correctly understood by non-experts. Vocabulary, associating technical terms with layman expressions, can help in increasing the readability of technical texts and their understanding. The purpose of our work is to build this kind of vocabulary. We propose to exploit the notion of reformulation following two methods: extraction of abbreviations and of reformulations with specific markers. The segments associated thanks to these methods are aligned with medical terminologies. Our results allow to cover over 9,000 medical terms and show precision of extractions between 0.24 and 0.98. The results and analyzed and compared with the existing work.

Keywords. Reformulation, Information Extraction, Medical Terminology, Layman Language, Patient Education

1. Introduction

Experts from medical area use sophisticated technical terms, which are usually non-understandable by patients [1]. This makes patients unable to make decisions, understand consequences of disorders and treatments, or even understand their disease. The situation is not improved by information increasingly available online [2]: patients remain often unable to understand it either. Find and understand new information implies solitary information retrieval process [3] when the patient is not accompanied by his medical doctor, with whom he may have verbal interactions and obtained needed information [4,5]. Indeed, with verbal interactions, it is possible to share knowledge and create common basis through reformulations, repetitions and clarifications [2].

Our purpose is to acquire vocabulary which associates specialized terms with layman expressions. We propose to exploit reformulations in texts written by experts or issued from collaborative media, in order to guarantee a better reliability of the extracted data. According to our hypothesis, (1) when experts reformulate terms, this indicates that the term is technical and conveys specialized meaning; (2) the reformulation act may allow to associate term with its reformulation; (3) reformulation is language phenomena spontaneously used in different kinds of texts.

In what follows, we present some existing work (section 2). We describe the material and methods used (section 3), and then present and discuss the results (section 4). We conclude with some directions for future work (section 5).

[1] Corresponding author, Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France. natalia.grabar@univ-lille3.fr

2. Related work

Reformulation occurs when someone is saying or writing again a given idea with different words [6], often in order to improve the understanding. Reformulation can be introduced by specific markers (*eg*, *let's say, That is*). Reformulated segments are not always semantically equivalent [7], but when they are it becomes possible to extract the paraphrases of technical terms. We can distinguish two main directions of the existing work: health literacy and automatic acquisition of paraphrases for medical terms. Health literacy is related to the understanding of medical information, its use and interpretation, and depends on two factors: environment (education, age, medical history...) and knowledge of patients, be it shared or not with the medical staff [8]. Several works have been dedicated to the acquisition of vocabularies associating terms with their paraphrases: exploitation of monolingual comparable corpora, in with morpho-syntactic patterns, similarity measures or n-grams allow to associate syntactic groups from both corpora [9,10]; exploitation of monolingual corpus for the acquisition of paraphrases for neoclassical compounds (*myocardial, desmorrhexia...*) thanks to the morphological analysis and segmentation of such terms, translation of their components in French (*myocardial: myo=muscle, cardia=heart*), and search of syntactic groups that contain these words (*heart muscle, muscle of the heart*) [11]. Let's also mention (1) the Consumer Health Vocabulary (CHV) in English [12], which a collaborative initiative, involving corpora processing, associative measures, and human annotators. The resource contains currently 141,213 unique terms; (2) an automatic translator of medical terms [13], best known through the resource MEDLINEplus, organized according to medical topics (tumors, obesity, etc). Let's also mention work on extraction of paraphrases, mainly applied to parallel and aligned corpora [14].

3. Material and Methods

We use the French part of the UMLS [15], and SNOMED International [16] (323,964 terms). We also use two corpora: (1) development corpus, issued from *masante.net* forum moderated by medical doctors. When users ask questions they are answered by the moderators. We use part with answers containing 6,139 answers (315,362 occ.); and (2) test corpus built with articles from the medical part of Wikipedia, which gives 18,434 articles (15,235,219 occ.). Finally, we use linguistic resources: 111 stopwords and morphological resources with 163,823 wordpairs (*eg* {*aorta, aortic*}) [17].

Our method is composed of four steps: (1) *pre-processing of corpora* with the syntactic analyzer Cordial [18]. Table 1 presents an example for *Vous devez les faire brûler par un gastroentérologue spécialisé, c'est-à-dire un proctologue (You must to make them burn by specialized gastroenterologist, that is a proctologist)*. The exploited fields are forms, lemmas, and syntactic information *type GS* and *Prop*; (2) *extraction of reformulations*, which is done with two approaches:

- for *abbreviations*, we use the raw corpus and extract two structures: *extended form (abbreviation)* and *abbreviation (extended form)*. For this, we implement an existing algorithm, which allows to associate each letter from abbreviation with a given word before of between parentheses [19]. Three situations are possible: *full*

when all letters from abbreviation are associated, *partial* when part of letters are associated, *null* when no letters are associated;

- for *reformulations with markers*, we exploit three markers *c'est-à-dire (That is), autrement dit (in other words), encore appelé (also called)* in the structure *concept marker reformulation*, like in *specialzed gastroenterologist, that is a proctologist*, where the underlined segments correspond to the source and target segments. It appears that source segments are better extracted with the *type GS* information, while target segments with the *Prop* information.

Table 1. An excerpt from syntactically tagged and analyzed text

form	lemma	POS	POSMT	GS	Type GS	Prop
Vous	vous	PPER2P	Pp2.pn	1	S	1
devez	devoir	VINDP2P	Vmip2p	2	V	1
les	le	PPER3P	Pp3.pa	3	C	2
faire	faire	VINF	Vmn--	4	D	2
brûler	brûler	VINF	Vmn--	5	V	3
par	par	PREP	Sp	8	F	3
un	un	DETIMS	Da-ms-i	8	F	3
gastroentérologue	gastroentérologue	NCMS	Ncms	8	F	3
spécialisé	spécialisé	ADJMS	Afpms	8	F	3
,	,	PCTFAIB	Ypw	-	-	3
c'	ce	PDS	Pd-..-	11	N	3
est	est	ADV	Rgp	-	p	3
-à	à	PREP	Sp	14	I	3
-dire	dire	VINF	Vmn--	14	I	3
un	un	DETIMS	Da-ms-i	8	F	3
proctologue	proctologue	NCMS	Ncms	16	D	3
.	.	PCTFORTE	Yps	-	-	-

(3) *alignment of the extracted segments with medical terminologies* has double objective: check the relevance of extractions and associate the extracted segments with medical terms. During the alignment, the extracted segments are normalized (accents, morphologically-related words), the stopwords are removed; and (4) *evaluation*, for which we build reference sets with two independent annotators for annotating source and target segments. For the evaluation of alignment with terminologies, we build a reference set from the development corpus. On the basis of the reference annotations, we can evaluate precision P, recall R and F-measure F of the extractions and alignments. Evaluation of extractions is performed with exact (boundaries of segments must be respected) and inexact (boundaries of segments can be inexact) versions.

4. Results and Discussion

The inter-annotator agreement [20] of extractions at the word level is 0.967 and 0.816, for alignments it is 0.208 and 0.714 for abbreviations and markers, respectively.

In the upper part of Table 2, we indicate number of extractions for each method: with abbreviations and the test corpus, we extract several candidates. With abbreviations, we observe three cases: correct extraction *{ESF; Editions Sociales Françaises}, {CD26; cluster de différenciation 26}*; partial correct extraction *{CHUM; Université Montréal}, {CHU; hôpital universitaire}*; partial incorrect extraction *{SEPP; plus}, {NFS; faire sang}*. Extractions with markers provide few duplicates

because reformulations are less controlled *(des canaux galactophores: qui fabriquent le lait de la femme, qui sécrètent le lait*). Evaluation of extractions indicates that abbreviations show 0.74 and 0.94 F-measure, while markers show 0.24 and 0.98 F-measure with exact (borders respected) and inexact versions, respectively. In the lower part of Table 2, we indicate number of alignments. Within aligned segments, we can observe 5 cases: full correct *{syndrome polyalgique idiopathique diffus; syndrome polyalgique idiopathique diffus.C0016053}*; morpho-syntactic variation *{troubles fonctionnels intestinaux; troubles gastrointestinaux fonctionnels/C0559031}*; partial *{semaines amenorrhee; amenorrhee/C0002453}*; compositional *cause/C0085978* and *pus/C0034161* for *cause de pus*; and incorrect (*{LCR; ph lcr/C0853364}*, *{liquide cerebro; regime liquide/C-F2300}*). The average F-measure for the two segments is 0.71 and 0.73 with abbreviations and markers, respectively.

Table 2. Number of extractions and alignments for each method.

	Development corpus		**Test corpus**	
	Abbrev.	**Markers**	**Abbrev.**	**Markers**
Extraction: nb occurrences	75	96	88,762	2,710
Extraction: nb types	42	96	8,106	2,710
Alignment: nb occurrences	75	96	88,762	2,757
Alignment: full alignments	11	5	154	42
Alignment: partial alignments	44	37	1,634	557
Alignment: not aligned	20	54	6,318	1,937

Table 3. Comparison with methods from the existing works.

Method	**Type of terms**	**Nb. extractions**	**Precision**
Abbreviations	abbreviations	42, 8,106	0.74/0.94
Markers	any type	96, 2,710	0.24/0.98
Definitions [11]	any type	1,028	0.52, 0.68
Morphology [11]	compounds	1,128	0.76, 0.86
N-grams [9]	morpho-syntactic	65, 82	0.67, 0.60
Syntactic groups [10]	morpho-syntactic	109	0.66
Abbreviations [18]	abbreviations	785	0.95

In Table 3, we propose a comparison with existing works: type of terms, number of extractions, precision (available for all cited works). We can see that our methods are efficient: they provide an important number of extractions with good precision.

5. Conclusion and Future Work

For the acquisition of vocabulary associating technical terms with layman expressions, we propose to exploit reformulation through two methods: extraction of abbreviations and their extended forms, and of reformulations introduced by markers. The methods are fixed on the development corpus and then applied to the test corpus. Exact precision is between 0.23 and 0.74, while inexact precision is between 0.68 and 0.98. The future work may study other markers. With the annotated corpora, we may apply supervised machine learning for making the extractions. The acquired vocabulary will be used for the simplification of medical and health documents.

Acknowledgement. This work has been performed as part of the *EQU (Éthique Qualité Urgence)* project funded by the call Thématique de l'établissement of université Lille 3.

References

[1] McCray A. Promoting health literacy. *J of Am Med Infor Ass* 2005;12:152-63.
[2] Zeng Q and Tse T. Exploring and developing consumer health vocabularies. *JAMIA* 2006;13:24-9.
[3] Boubé N and Tricot A. *Qu'est-ce-que rechercher de l'information ?: état de l'art*. Presses ENSSIB 2010.
[4] Vergely P, Condamines A, Fabre C, et al. Analyse linguistique des interactions patient/médecin. *Actes éducatifs de soins* 2009;92(5).
[5] Zielstorff RD. Controlled vocabularies for consumer health. *Journal of Biomedical Informatics* 2003;36(4-5):326-3.
[6] Le Bot MC, Schuwer M, and Richard E (dir.). *La reformulation : Marqueurs linguistiques – Stratégies énonciatives*. Rivages linguistiques, Rennes, 2008.
[7] Gulich E and Kotschi T. Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française* 1983;5:305-51.
[8] Zeng QT, Kim E, Crowell J, and Tse T. A text corpora-based estimation of the familiarity of health terminology. In: ISBMDA 2006, 184-92.
[9] Deléger L and Zweigenbaum P. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In: AMIA 2008, 146-50.
[10] Cartoni B and Deléger L. Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes. In: TALN, 2011.
[11] Grabar N and Hamon T. Extraction automatique de paraphrases grand public pour les termes médicaux. In: TALN 2015, Caen, France. 14 p.
[12] Zeng QT, Tse T, Divita G, et al. Exploring lexical forms: first-generation consumer health vocabularies. In: AMIA 2006, 1155
[13] McCray A, Loane R, Browne A, and Bangalore A. Terminology issues in user access to web-based medical information. In: AMIA Symposium 1999, 107-7.
[14] Androutsopoulos I and Malakasiotis P. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research,* 2010:38, 135-187
[15] Lindberg D, Humphreys B, and McCray A. The unified medical language system. *Methods Inf Med* 1993;32(4):281-91.
[16] Côté RA, Rothwell DJ, Palotay JL, Beckett RS, and Brochu L. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield, 1993.
[17] Grabar N and Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP* 2000:310-4.
[18] Laurent D, Nègre S, and Séguéla P. Apport des cooccurrences à la correction et à l'analyse syntaxique. In: TALN, 2009.
[19] Schwartz AS and Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. In: Pacific Symposium on Biocomputing, 2003:451-6.
[20] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20(1):37-46.