Identification of Drug-Related Medical Conditions in Social Media

François Morlane-Hondère Cyril Grouin Pierre Zweigenbaum
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
{francois.morlane-hondere,cyril.grouin,pz}@limsi.fr

Abstract

Monitoring social media has been shown to be an interesting approach for the early detection of drug adverse effects. In this paper, we describe a system which extracts medical entities in French drug reviews written by users. We focus on the identification of medical conditions, which is based on the concept of post-coordination: we first extract minimal medical-related entities (*pain*, *stomach*) then we combine them to identify complex ones (*It was the worst [pain I ever felt in my stomach]*). These two steps are respectively performed by two classifiers, the first being based on Conditional Random Fields and the second one on Support Vector Machines. The overall results of the minimal entity classifier are the following: P=0.926; R=0.849; F1=0.886. A thourough analysis of the feature set shows that, when combined with word lemmas, clusters generated by word2vec are the most valuable features. When trained on the output of the first classifier, the second classifier's performances are the following: p=0.683;r=0.956;f1=0.797. The addition of post-processing rules did not add any significant global improvement but was found to modify the precision/recall ratio.

Keywords: Pharmacovigilance; Natural Language Processing; Named Entity Recognition

1. Introduction

Most modern drugs and medicines are known to cause adverse reactions. Some of them are detected during premarketing clinical trial and documented while others are not. Thus, it is crucial to monitor the post-marketing phase of a drug to determine whether its risk/benefit ratio is still positive or not. This activity is called pharmacovigilance.

Today, adverse drug reactions are reported spontaneously by patients and professionals via online forms or phone calls to regulating authorities like the Food and Drug Administration (FDA) in the United States or the Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM) in France. However, studies and reports have shown the relative inefficiency of this reporting method, which is mainly due to the lack of awareness of patients and professionals (Alshakka et al., 2013; Académie Nationale de Pharmacie, 2014).

Improving the efficiency of the way drug-related problems are identified is critical given the human and economic costs of adverse drug reactions. Recent years have seen the birth of what can be called *computer-assisted pharmacovigilance* or *e-pharmacovigilance* (Neubert et al., 2013). This consists in using text mining to extract pharmacovigilance-related information from sources such as biomedical articles, electronic health records or drug labels (Harpaz et al., 2014; Shang et al., 2014).

Social media have also been used for that purpose. Characteristics of user-generated content are valuable for computer-assisted pharmacovigilance in that it is massive, continuously generated and easy to access. It is also challenging in that the informal writing style that defines user-generated content has been shown to deteriorate the performance of traditional NLP tools (Gimpel et al., 2011). A recent review describes twenty-two studies that aim at extracting adverse drug reactions from tweets and various health forums (Sarker et al., 2015). Twenty-one of these studies are based on English data.

This paper follows these previous studies by searching for medical conditions in user-generated French drug reviews. The concept of medical condition subsumes both pathologies that are caused by a medicine ('side effects') and pathologies that are the reason for the medication ('indications'). Differentiating side effects from indications has been found to be complex (Nikfarjam et al., 2015). We will tackle this problem in future work.

We treated the identification of medical conditions as a two-step classification task. A first classifier is used to identify medically-related named entities, a second one is used to identify relations between the entities that form a complex medical condition. The assumption is that it is easier to identify *cramps* and *legs* in *I have been getting cramps at night in my legs* then a dependency relation between them than to identify the whole sequence *cramps at night in my legs*. Both classifiers are trained on a manually-annotated 1,200 reviews corpus that will be made available soon.

2. Material and methods

2.1. Corpus

2.1.1. Presentation

Our corpus is composed of 12,440 French drug reviews extracted from the website `meamedica.fr`. The main reason why we targeted drug reviews is because nearly 80% of them explicitly mention adverse reactions, which is much higher than in traditional forum posts (24%) (O'Connor et al., 2014). Drug reviews are different from messages posted in traditional health forums in that they do not appear in a conversation thread (users can give feedbacks on another user's review but this feature is hardly used). Nevertheless, as forums posts, they are user-generated and written in an informal style. As can be seen in Figure 1, users are invited to provide personal information as well as star ratings for parameters such as drug effectiveness, ease of use, presence/severity of adverse reactions and general satisfaction. In this study, only the textual part of the review has been extracted. The length ranges from a few words mentioning the adverse reactions ('Damaged thyroid!') or the overall impression about the drug ('Not great.') to narratives longer than a thousand words, the average message length being 89 words.

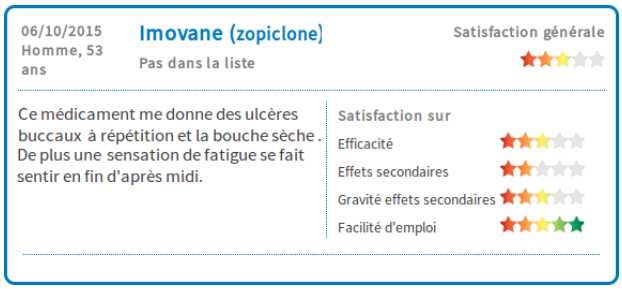

Figure 1: A drug review posted on `meamedica.fr`

2.1.2. Annotations

We chose to annotate 1,200 randomly selected reviews, which is slightly higher than the amount of annotated data used in similar studies (Sarker et al., 2015). Annotations were made using the BRAT annotation tool (Stenetorp et al., 2012). A set of 100 reviews has been annotated in duplicate by two humans (the first two authors of the paper). We computed an inter-annotator agreement between annotations produced by both annotators. The high agreement (kappa = 0.825) suggests that this task can be reliably performed by a single annotator. As a consequence, the remaining 1,100 reviews were annotated by only one human. Figure 2 presents the corpus production process.

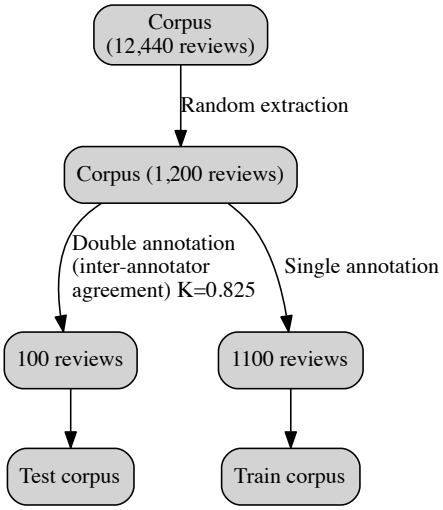

Figure 2: Corpus constitution

The annotation process followed a two-fold strategy. First, we annotated in the corpus all the occurrences of 16 semantic classes inspired from the UMLS Metathesaurus (Bodenreider, 2004), as presented in the guidelines we defined. These classes can be broken into three groups:

- **drug-related information:** CHEMICALS OR DRUGS, CONCENTRATION, DOSAGE, MODE (galenic formulation);
- **clinical information:** ANATOMY (body parts, including fluids and tissues), DISORDERS (diseases), FUNCTION (functions of the human body like breathing or hearing), GENES/PROTEINS, MEDICAL PROCEDURE, SIGN OR SYMPTOM;

class	train	test	ratio
SIGNSYMPTOM	3862	310	12.5
ANATOMY	2085	137	15.2
CHEMICAL	2085	155	13.5
DURATION	1318	99	13.3
DISORDERS	1100	88	12.5
FUNCTION	947	75	12.6
PROCEDURE	739	64	11.5
JOB	670	61	11.0
DOSAGE	584	59	9.9
TIME	439	60	7.3
DATE	393	35	11.2
FREQUENCY	332	29	11.4
MODE	311	28	11.1
CONCENTRATION	114	9	12.7
WEIGHT	86	5	17.2
GENE	19	3	6.3

Table 1: Number of annotated entities of each semantic classes in the train and test corpus

- **complementary information:** JOB, WEIGHT and temporal information (DATE, TIME, DURATION, FREQUENCY).

The number of entities annotated for each class is shown in Table 1. Although some of those classes are not directly related to medical conditions, dosage, concentration, mode, weight and time-related information are very important in that they can be potential indicators of drug misuse.

To make the manual annotation process easier, we used a pre-annotation method consisting in bootstrapping our classifier (section 2.3.) with the 100 first annotated messages, then pre-annotating the 100 next ones, which were in turn, after manual validation, added to the training set, etc. This step was performed until the quality of the pre-annotation was found acceptable.

Then, we annotated *expansion relations*, which are defined as syntactico-semantic dependencies between pathologies (DISORDERS/SIGN OR SYMPTOM) and their objects or locations (ANATOMY, FUNCTION). Exemples of such relations are given in Figure 3: *baisse* ('lapse') is linked to *concentration* ('focus'), *coloration* ('colouration') to *urine* ('urine'), *douleur* ('pain') to *dos* ('back'). 2,426 expansion relations have been annotated.

2.2. Pre-processing

One of the main problems with user-generated content is that it is prone to spelling errors and erroneous syntactic structures, which negatively affect the performance of NLP tools such as parsers. However, it has been shown that simple transformations could increase significantly the performances of parsers on forum posts (Foster, 2010).

Thus, in order to improve the parsing of our corpus, we extracted the most frequent unknown words in a tagged 17.5 million word health forum[1] corpus—the 'Atoute corpus'—and wrote a set of 80 rules to replace them with their normalized forms. Most of the rules are designed

[1] `http://www.atoute.org/n/forum/`

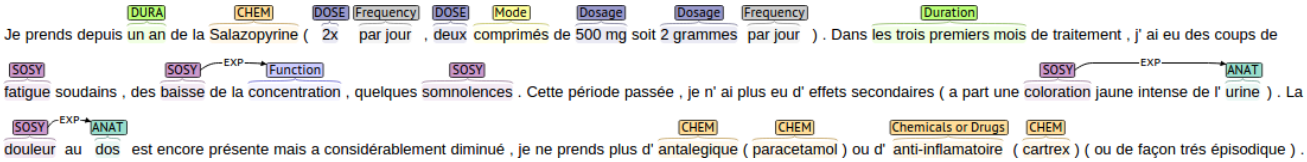

Figure 3: Excerpt of an annotated message.

to target function words (*ca*→*ça*, *ki*→*qui*, *pr*→*pour*) but some lexical words are also replaced (*pb*→*problème*, *ad*→*antidépresseur*, *pds*→*prise de sang*). This corpus was then parsed using the Talismane parser (Urieli, 2013).

2.3. Entity extraction

Because of their good performance for sequence labelling, Conditional Random Fields (CRFs) (Lafferty et al., 2001) have been widely used in named entity recognition tasks. In this study, we used the CRF-based Wapiti toolkit (Lavergne et al., 2010) trained on 1,100 annotated messages (the 100 remaining ones being used as a test set).
The classification is based on the following sets of features (features followed by an asterisk have been used in conjunction with all other features):

- lemmas:
  - current*, previous* and next* lemma alone;
  - lemma bigrams.
- current*, previous and next part of speech;
- syntactic function*;
- token's length* (in characters);
- the presence of a number in the token*;
- token's case (*MM* = all letter uppercase, *mm* = all letters lowercase, *Mm* = first letter uppercase);
- token's Soundex code. Soundex is an algorithm that indexes words based on their phonetic properties. Thus, words that sound similar have similar values. This can be used to identify words' variations or misspellings (Kondrak and Dorr, 2004). For example, *arythmie* and *arytmie* share the Soundex code A635;
- lemma's three initial and final characters (the lemma must be at least five characters long). If no lemma could be identified, this applies to the token;
- the presence of the current and the three previous lemmas in:
  - a hand-picked list of cause/consequence markers (*effect*, *result*, *responsible*);
  - a semi-automatically extracted list of drug-introducing verbs like *take* or *swallow* (Morlane-Hondère et al., 2015).
- the presence of the lemma in a hand-picked list of temporal markers (*often*, *daily*, *morning*);
- the lemma's clusters identifiers*. Word clustering consists in making n groups of words according to their distribution in a corpus, relying on the assumption that words that occur in similar contexts tend to have similar meanings (the 'distributional hypothesis' (Harris, 1954)). Word clusters were generated using the software word2vec (Mikolov et al., 2013) on the lemmatized and lower-cased Atoute corpus. We considered different levels of granularity by generating four models with respectively 100, 200, 400 and 800 clusters (and 3 as window size). As shown by Nikfarjam et al. (2015), this method allows the grouping of similar entities like drugs or symptoms. Table 2 shows some of the semantic groupings generated by word2vec (with 400 clusters) and the label which could be attributed to them. This method is also useful to retrieve spelling variations or misspellings. For example, we found 19 variations of the word *gynécologue* in the same cluster (*gynécolgue*, *gynécoloque*, *gynéco*, *gynécho*, *génico*, *gyné*, *gygy*...);
- the token appears in a drug list composed of: the 8,691 French entries in the UMLS from the *Pharmacologic substance* semantic type, 9064 terms found in the list of generic drugs[2] provided by the French pharmacovigilance agency (ANSM), 10,870 drug names extracted from EurekaSanté, a French online dictionary of drugs[3] providing general information on drugs, the list of the 100 more prescribed drugs in France[4];
- the current, two previous and two next lemma's UMLS semantic class.

2.4. Relation extraction

We then used support vector machines (SVM) as implemented in the LibSVM software (Chang and Lin, 2011) to identify expansion relations between the different components of a medical condition. We generated all possibles expansion relations between the DISORDERS, SIGN OR SYMPTOM, ANATOMY and FUNCTION entities in a 7 words window (this value was found to be best in terms of precision/recall). The goal of the SVM is to decide, for each

[2] http://ansm.sante.fr/var/ansm_site/storage/original/text/97b3c42da571c69da1e837f759076675.txt

[3] http://www.eurekasante.fr/medicaments/alphabetique.html

[4] http://www.doctissimo.fr/asp/medicaments/les-medicaments-les-plus-prescrits.htm

cluster label	size	members
drug names	107	*melodia*, *utrogestan*, *zoely*, *evepar*, *desobel*, *triella*…
medical exams	133	*cytoponction*, *électro-encéphalogramme*, *ostéodensitométrie*…
medical intervention	109	*hépatectomie*, *mastectomie*, *pontage*, *laparoscopie*, *lithotricie* [sic]…
quantities/measures	188	*2ml*, *60g*, *microgrammes*, *225mgr*, *68kg*, *1m64*…
semantic field of pregnancy	51	*amniocentèse*, *extra-uterine*, *fiv*, *grossese*, *insémination*…
food	163	*biscuit*, *steak*, *tomate*, *chocolat*, *mayonnaise*…

Table 2: Examples of semantic groupings generated by word2vec

pair of entities, whether or not it is linked by an *expansion* relation.
The model was trained on the relations manually annotated in the 1,100 messages used by the first classifier. This second classifier was trained using the following features:

- length of the span between the two linked entities (in number of words);
- parts of speech of the linked entities;
- parts of speech of the words in the span;
- annotations of the linked entities;
- parts of speech of the words in the span;
- syntactic functions of the linked entities.

Post-processing rules are used to refine the results provided by the SVM. For exemple, one of the rules automatically assigns the value FALSE to relations between two words separated by punctuation marks like periods, colons, semicolons, etc.

3. Results

In order to evaluate the features used in the CRF, we trained a series of models based on two feature sets: lemma unigrams and bigrams—which were used as a baseline—and each one of the remaining 13 features. By measuring the predictive power of each feature in a minimal model, we were able to identify the features which may have a positive or negative impact on the overall performance. The overall f1 scores obtained by each of these models are given—sorted in ascending order—in table 3.
We then measured the ability of the SVM classifier to identify the 160 expansion relations of the test set. In a first setting, the SVM model is applied on the output of the CRF. In the second setting, we used the manual annotations as an input for the SVM. The application of the model on an imperfect automatically-generated annotation and a manual one allows us to measure the performance loss that can be attributed to the quality of the input or to the SVM itself. The results are provided in Table 5.

4. Discussion

4.1. Entity extraction

We can see in Table 3 that only the combination of drug-introducing verbs and lemmas gives a lower score than lemmas alone. Temporal markers have no influence on the baseline. Word clusters are the feature which give the best

feature	f1	diff
lemma	0.789	–
" + drug-introducing verbs	0.786	– 0.003
" + temporal markers	0.789	0.000
" + part of speech	0.792	+ 0.003
" + consequence markers	0.793	+ 0.004
" + presence of a number	0.794	+ 0.005
" + case	0.799	+ 0.010
" + syntactic function	0.805	+ 0.016
" + token length	0.805	+ 0.016
" + semantic class	0.809	+ 0.020
" + drug list	0.811	+ 0.022
" + Soundex code	0.827	+ 0.038
" + prefix/suffix	0.839	+ 0.050
" + clusters	0.852	+ 0.063
all	0.881	+ 0.092
" – drug-introducing verbs	0.886	+ 0.097

Table 3: Predictive power of the features used in the CRF

class	p	r	f1
JOB	1.000	0.983	0.991
MODE	1.000	0.964	0.981
ANATOMY	0.962	0.941	0.952
CHEMICAL	0.979	0.922	0.950
DURATION	0.956	0.888	0.921
TIME	0.912	0.866	0.888
SIGNSYMPTOM	0.890	0.838	0.863
PROCEDURE	0.925	0.781	0.847
DOSAGE	0.844	0.830	0.837
DATE	0.875	0.800	0.835
FUNCTION	0.932	0.733	0.820
DISORDERS	0.915	0.738	0.817
FREQUENCY	0.875	0.724	0.792
WEIGHT	1.000	0.600	0.750
CONCENTRATION	0.500	0.333	0.400
GENE	0.000	0.000	0.000
overall	0.926	0.849	0.886

Table 4: Precision, recall and f1 score obtained for each class

input	method	p	r	f1
CRF	SVM alone	0.683	0.956	0.797
output	SVM + post-proc.	0.746	0.881	0.808
Gold	SVM alone	0.724	0.969	0.829
annotation	SVM + post-proc.	0.770	0.900	0.830

Table 5: Precision, recall and f1 score obtained by the SVM with and without post-processing

results in combination with lemmas, which supports the results obtained by Nikfarjam et al. (2015).

The best score is achieved by a model made of all the features but drug-introducing verbs. These verbs were semi-automatically extracted from the Atoute corpus in a previous study (Morlane-Hondère et al., 2015) on the basis that they are prone to be used when users write about their medication: *il m'a prescrit du Roaccutane 20 mg* ('I have been prescribed Roaccutane 20 mg'), *je suis passé au Cymbalta* ('I switched to Cymbalta'). We made the assumption that this feature would benefit to the extraction of CHEMICAL entities—especially the misspelled ones— but the addition of this feature decreases the efficiency of the classifier, both at the global scale and for the CHEMICAL class. One explanation is that some of these verbs (*prendre* 'take', *donner* 'give', etc.) are too polysemous and should be disambiguated for the feature to be discriminative.

Although other features like parts of speech or consequence markers bring no significant improvement, we decided to keep them in the final model as their predictive power may increase in combination with other features.

Best performing model's performances for each class are given in table 4. Classes are sorted according to the decreasing f1 score. Except from the three classes with less than 10 occurrences in the test corpus—WEIGHT, CONCENTRATION and GENE—all the classes have a f1 score approaching or higher than 0.800.

4.2. Expansion relation extraction

Results in Table 5 show that the f1 score difference between the model trained on the CRF output and the model trained on the gold annotation is 0.032. The difference is smaller (0.022) between the post-processed results. The post-processing appears to have only a small impact on the f1 scores of both models. However, their precision/recall ratio increases from 0.71 to 0.85 (CRF input) and to 0.75 to 0.85 (gold annotation). This means that the post-processing patterns tend to balance the precision/recall ratio by filtering out a nearly equal amount of false and true positives.

5. Conclusion

We developed a classification system to extract medical conditions and complementary medical-related information in French drug reviews. It was trained on a soon to be available manually-annotated French drug reviews corpus.

This system followed a two-step procedure. The first step consisted in using a CRF classifier trained on manually-annotated data to identify various medical entities. The second one was used to identify relations between the entities extracted by the CRF to extract complex medical conditions. The good results of the CRF for the four classes used in the second step—ANATOMY, SIGNSYMPTOM, FUNCTION and DISORDERS—allowed the SVM to perform quite well (0.033 below the performances of the same model with gold annotation as input). We also showed that the use of post-processing patterns would not lead to a substantial increase but would be useful to balance the precision/recall ratio.

Acknowledgement

This work was supported by the ANSM (French National Agency for Medicines and Health Products Safety) through the Vigi4MED project under grant #ANSM-2013-S-060.

6. References

Académie Nationale de Pharmacie. (2014). Rapport et recommandation sur la facilitation de la notification directe d'effets indésirables par les patients. `http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf` [Last visited on 10/19/2015].

Alshakka, M. A., Ibrahim, M. I. M., and Hassali, M. A. A. (2013). Do health professionals have positive perception towards consumer reporting of adverse drug reactions? *J Clin Diagn Res*, 7(10):2181–5.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*, 2:27:1–27:27.

Foster, J. (2010). "cba to check the spelling" investigating parser performance on discussion forum posts. In *Proc of HLT*, pages 381–384, Los angeles, CA.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proc of HLT*, pages 42–47, Portland, OR.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., and Shah, N. (2014). Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Safety*, 37(10):777–790.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Kondrak, G. and Dorr, B. (2004). Identification of confusable drug names: A new approach and evaluation methodology. In *Proc of Coling*, Geneva, Switzerland.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proc of ACL*, pages 504–513, Uppsala, Sweden.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2015). Étude des verbes introducteurs de noms de médicaments dans les forums de santé. In *Actes de TALN*, pages 337–343, Caen, France.

Neubert, A., Dormann, H., Prokosch, H.-U., Bürkle, T., Rascher, W., Sojer, R., Brune, K., and Criegee-Rieck, M. (2013). E-pharmacovigilance: development and implementation of a computable knowledge base to identify

adverse drug reactions. *Brit J Clin Pharmacol*, 76:69–77.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*, 22(3):671–681.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K., and Gonzalez, G. (2014). Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. In *Proc of AMIA*, Washington, DC.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *J Biomed Inform*, 54:202 – 212.

Shang, N., Xu, H., Rindflesch, T. C., and Cohen, T. (2014). Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J Biomed Inform*, 52:293–310.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In *Proc of EACL Demo*, pages 102–107, Avignon, France.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

A Annotation guidelines

In this section, we present the annotation guidelines we used while producing the corpus used in this study. The guidelines is composed of 16 categories which are similar to semantic types from the UMLS.

A1. Annotation rules

We defined the following set of annotation rules: entities are annotated according to their semantic information; annotations are made on single words as much as possible (except for portions giving useful information, e.g., *wisdom teeth* vs. *teeth*); determiners and prepositions are not annotated, except if they give information (e.g., *from time to time* and not only *time to time*); annotation of all entities, whatever the spelling (e.g., *parcetamol* instead of *paracetamol*).

A2. Categories

We defined three main kinds of categories: (i) drug-related information, (ii) clinical information, and (iii) additional information such as social data and temporal data.

A2.1. Drug-related information

Chemical or Drug Drug name, active substance, pharmacological class (including acronyms).

(1) Actuellement je prends chem Abilify depuis 2 jours et j'ai déjà perdue 1kg.

(2) Certains chem ADO [5] ont des effets secondaires, propres à chacun..

[5] ADO: anti-diabétique oral (oral anti-diabetic).

Concentration Concentration of a drug, generally associated with treatment presented as tablet. Concentration must be distinguished from dosage.

(3) J'ai essayé 2 cachets de parcetamol conc 500/50 à la codéïne.

Dosage Dosage of a treatment; galenic form must not be annotated as "dosage" but as a "mode".

(4) Mon médecin m'as prescrit du Mediator pour faire chuter mon taux de triglycerine a raison de dose 3 comprimés / jour.

Mode Galenic form of the drug *(tablet, gellule, syrup)*.

(5) A chaque prise de l'mode ampoule, je suis tombée sans raison quelques heures après, jambe qui se dérobe.

(6) Mon médecin m'as prescrit du Mediator pour faire chuter mon taux de triglycerine a raison de 3 mode comprimés / jour.

A2.2. Clinical information

Anatomy All body parts, including fluids and tissues. *Remark: annotations are made on both nouns and adjectives.*

(7) Cependant j'ai mon anat ventre qui gargouille beaucoup et des anat selles liquides.

(8) Effets secondaires : 24/7 léger mal de anat tête avec de temps en temps de fortes poussées surtout du côté droit du anat cerveau.

(9) Mais depuis j'ai eut un cancer à la anat thyroïde.

Genes Proteins Proteins, lipids, nucleic acids, genes.

(10) Mon médecin m'as prescrit du Mediator pour faire chuter mon taux de prot triglycerine a raison de 3 comprimés / jour.

Biological Process or Function Natural process or state, or resulting from an activity.

(11) C'est un médicament pour diabétique et non pour func maigrir.

(12) J'ai du mal à func respirer la nuit (angoisses ???)

(13) J'avais régulièrement mal à la tête et je func voyais moins bien.

(14) Mon médecin m'a prescrit du Kestin, qui me mettait dans un état second (fatigue, incapacité à me func concentrer...).

Disorders Concern both names of illness than name of patients suffering of this illness since it is possible to infer the name of the illness from this adjective.

(15) C'est un médicament pour diso diabétique et non pour maigrir.

(16) Dernière prise, crise de diso folie, sensation de diso mort imminente, cauchemars constant.

(17) Mais depuis j'ai eut un diso cancer à la thyroïde.

(18) Prise de Lariam en prophylaxie contre le diso paludisme.

Sign or Symptom Observable manifestation of illness, based on a clinical observation. Familiar terms (sentence 20) must be annotated since they correspond to problems experienced by patients.

(19) A chaque prise de l'ampoule, je suis sosy tombée sans raison quelques heures après, jambe qui se sosy dérobe.

(20) Cependant j'ai mon ventre qui sosy gargouille beaucoup et des selles sosy liquides.

(21) J'avais régulièrement sosy mal à la tête et je voyais sosy moins bien.

Medical Procedure All medical procedure, including diagnoses, procedures, exams and treatment methods.

(22) Malgré le proc régime à la dernière proc prise de sang j'avais encore 3.50g/l.

(23) Prise de Lariam en proc prophylaxie contre le paludisme.

A2.3. Additional information

This last set of categories gives useful information to complete clinical data.

Job Professional activity, most of the time a medical activity, concerning the professional mentioned by the user.

(24) Mon job médecin m'a prescrit du Kestin, qui me mettait dans un état second (fatigue, incapacité à me concentrer...).

Weight Total weight, won or lost.

(25) Actuellement je prends Abilify depuis 2 jours et j'ai déjà perdue wght 1kg.

Date Absolute or relative date, the most specific.

(26) C'est en date mars 2009 que j'ai été opérée à coeur ouvert pour changer 2 valves contre des valves mécaniques.

Duration Duration of treatment or illness. The temporal trigger word will generally not be included *(since, during)*.

(27) A chaque prise de l'ampoule, je suis tombée sans raison dura quelques heures après, jambe qui se dérobe.

Frequency Frequency of treatment or problem.

(28) Effets secondaires : freq 24/7 léger mal de tête avec freq de temps en temps de fortes poussées surtout du côté droit du cerveau.

(29) Mon médecin m'as prescrit du Mediator pour faire chuter mon taux de triglycerine a raison de 3 comprimés freq / jour.

Time Part of the day *(morning, evening, night)*.

(30) J'ai du mal à respirer la time nuit (angoisses ???)