Clinical Case Reports for NLP

Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon

▶ **To cite this version:**

Cyril Grouin, Natalia Grabar, Vincent Claveau, Thierry Hamon. Clinical Case Reports for NLP. BioNLP 2019 - 18th ACL Workshop on Biomedical Natural Language Processing, Aug 2019, Florence, Italy. pp.273-282, 10.18653/v1/W19-5029 . hal-02371243

HAL Id: hal-02371243
https://hal.science/hal-02371243

Submitted on 4 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

Clinical Case Reports for NLP

Cyril Grouin
LIMSI, CNRS, Université Paris Saclay
Campus universitaire d'Orsay
91405 Orsay cedex, France
cyril.grouin@limsi.fr

Natalia Grabar
STL, CNRS, Université de Lille
Domaine du Pont-de-Bois
59653 Villeneuve-d'Ascq cedex, France
natalia.grabar@univ-lille.fr

Vincent Claveau
IRISA, CNRS
Czampus universitaire de Beaulieu
35042 Rennes cedex, France
vincent.claveau@irisa.fr

Thierry Hamon
LIMSI, CNRS, Université Paris Saclay
Université Paris 13
99 avenue Jean-Baptiste Clément
93430 Villetaneuse, France
thierry.hamon@limsi.fr

Abstract

Textual data are useful to access expert information. Since the texts are representative of distinct language uses, it is necessary to build specific corpora in order to be able to design suitable NLP tools. In some domains, such as medical domain, it may be complicated to access the representative textual data and their semantic annotations, while there exists a real need for providing efficient tools and methods. In this paper, we present a corpus of 717 clinical cases written in French. We manually annotated this corpus into four general categories (age, gender, outcome, and origin) for a total number of 2,835 annotations. The values of age, gender, and outcome have been normalized. We also manually annotated a subset of 70 files into 27 fine-grained categories, for a total number of 5,198 annotations. In addition, we present a few basic experiments made on those annotations in order to highlight their usefulness.

1 Introduction

In Natural Language Processing (NLP), texts are useful to access information, especially expert information. Nevertheless, the linguistic diversity (type of narratives, common or specialized vocabulary, regular or complex syntactic structures, etc.) requires robust tools to access the information present in those texts. In order to build suitable NLP-based tools, to model linguistic elements (machine-learning, word-embeddings), or to produce gold standards for evaluating automatic systems, texts are needed (Nadkarni et al., 2011). However, due to privacy and ethical reasons, documents from specialized domains (e.g., clinical notes or justice decisions) are not easily accessible unless authorization (Chapman et al., 2011). When such data exist for the research, they are generally limited to English language, such as the MIMIC-III database (Johnson et al., 2016) and derived corpora. For French language, the Quaero medical corpus (Névéol et al., 2014) is composed of a limited number of documents (13 documents from the European Medicines Agency, 25 documents from the European Patent Organization) or very short documents (2,500 Medline titles).

In order to make available documents concerned by privacy issues, de-identification techniques have been widely used to replace nominative data by plausible information (Meystre et al., 2010; Kayaalp, 2017). Despite the recent improvements of these techniques, especially based on artificial neural networks (Dernoncourt et al., 2017), one can not assure that all nominative data have been removed and humans must further check those documents. Another solution relies on the production of synthetic data (Lohr et al., 2018). Originally, they were generated and used to train OCR systems for handwriting recognition (Doermann and Yao, 1995). They are now used when original data are missing or to provide more data, despite their artificial character (Eger et al., 2019). Besides, whether the texts are de-identified or artificially generated, their linguistic specificity will have an impact on further designed NLP rule-based and statistically-based approaches.

In this paper, we present the semantic annotations we made on a corpus of clinical cases written in French by domain experts. Since this corpus is composed of already published and freely accessible clinical cases, our aim is to make this annotated corpus available for the research. In order to present the usefulness of those annotations, we present a few basic experiments we made.

Proceedings of the BioNLP 2019 workshop, pages 273–282
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

2 Corpus and annotation guidelines

2.1 Corpus

In the clinical domain, in order to overcome the privacy and ethical issues when working on electronic health records, one solution consists in using clinical case reports. Indeed, it is quite common to find freely available publications from scientific journals which report in clinical cases of real de-identified or fake patients. Such clinical cases are usually published and discussed to improve medical knowledge (Atkinson, 1992) of colleagues and medical students. One may find scientific journals specifically dedicated to case reports, such as the *Journal of Medical Case Reports* launched in 2006 (Rison et al., 2017). Clinical cases consist of a detailed and hierarchically structured description of history, signs and symptoms, diseases, tests, treatments, follow-up and outcome of a given patient or of a cohort of patients (Rison, 2013). As pinpointed by Lysanets et al. (2017), clinical cases are composed of linguistic particularities which constitute a specific genre of medical texts: active voice sentences, past simple tense, personal pronouns, and modal verbs. Beyond this warning, they represent both an available and useful clinical content, especially for the NLP community for which the access to EHRs is becoming harder and harder.

We assume that this new orientation to tackle the medical data accessibility problem may become popular in the years to come within the biomedical domain. Let's for instance mention the work by Satomura and Amaral (1992), which produced back in 90's an automatic system designed for the indexing of clinical cases with ICD-9 codes. These clinical cases written in English have been extracted from the *New England Journal of Medicine* and permitted the researchers to develop their NLP system and to test it. More recently, Gurulingappa et al. (2012) produced a benchmark corpus composed of 3,000 clinical case reports in English, which has been then annotated into several categories (drug, dosage, and adverse effects), and relationships among them in order to provide mentions of adverse drug reactions.

The corpus we present in this work is composed of 717 clinical case reports written in French (see table 1 for general statistics). These cases have been previously published and are freely accessible. The cases from scientific literature often go with their discussion and keywords. In this work, we only focus on the clinical case description. This set has been manually annotated with general and fine-grained information, which is described in the two following sections. This corpus is part of a larger and yet growing corpus, which currently contains over 4,100 clinical cases (Grabar et al., 2018).

Element	Number
Documents	717
Sentences	1,124
Words (occurrences)	26,787
Words (forms)	5,030

Table 1: General statistics on the corpus annotated in this work

2.2 Annotations of general information

We considered four general categories of information for the annotation. They are related to demographic data (age and gender) and to medical data (the starting medical problem or origin and the outcome). Most of the clinical cases describe the clinical events of one patient. Yet, some clinical cases may be dedicated to the description of several patients, in which case, all relevant information are annotated for each patient. For this reason, the total number of annotations may be higher than the number of clinical cases. For three out of four categories, the values are normalized and taken from finite sets:

- Age $\in \mathbb{N}$: numerical value rounded in years; age in letters is converted into numerical value;
- Gender $\in \{$ feminine, masculine $\}$;
- Outcome $\in \{$ recovery, improvement, stable condition, worsening, death $\}$.

Besides, when several ages are given for the same patient, only the age at the moment of the main clinical event is considered. For the category Origin, the values correspond to text spans describing the initial medical problem.

Two scientists with a biomedical computer science background created the annotations independently, and then elaborated consensual annotations. Hence, all spans of text providing the expected information were annotated. For the category origin, the most inclusive text spans have been chosen.

2.3 Annotations of fine-grained information

The corpus has also been enriched with fine-grained annotations of entities concerning physiology, surgery, diseases, drugs, temporal data, lab and exam results. The annotations are based on the semantic types from the UMLS (Lindberg et al., 1993), on existing annotation guidelines such as the I2B2 NLP Challenges (Uzuner et al., 2010, 2011), and on medical entities from our corpus. We provide those annotations as a basis for several NLP tasks such as information extraction or automatic classification based on clinical entities.

In this section, we present the guidelines we defined. For each category, we give a definition and a few examples from the corpus.

2.3.1 Physiology

Body measurements: weight *(71.8 kg)*, size *(165 cm)*, and body surface area *(1.81 m^2)*

Vital signs: temperature *(38.2 ^oC)*, and physiological liquid mentions *(blood, urine)*

Biology: anatomical parts *(left lung, thyroid)*, localization of procedures or diseases *(arterial, pulmonary)*, and biological functions *(pregnancy, pulse)*

2.3.2 Surgery

These categories are related to the surgery:

Medical speciality including the types of medical units *(oncology, surgical care units)*.

Tests including names of tested elements *(radiography, biological check-up, blood pressure)*

Surgical treatments: treatments done by physicians *(chemotherapy, resection)*

Surgical approach: access used by the physician *(apical access)*

Medical devices used by patients or by physicians *(drainage, mask, sensor)*

2.3.3 Diseases

We considered four types of disease-related information:

Pathology: mentions of diseases or diseased condition *(acute lymphoblastic leukemia, tumor)*

Signs or symptoms which are not chronic diseases *(cough, fever, headache, hypertension)*

Biological organism: bacteria and infectious organisms *(escherichia coli, group B streptococcus)*

Nature: indication of quality (qualifying adjectives, grade) for diseases, signs and symptoms *(pT2 G1 carcinoma, benign cyst)*

2.3.4 Drugs

Pharmaceutical class or family of drugs *(antibiotic, anticoagulant, anti-vitamin K)*

Substance: commercial and generic drug names or generic substance *(acetaminophen, ferrous sulphate)*

Concentration of molecules in drugs *(10%, 5 mg/ml)*

Mode of administration *(intravenous, oral route, by nebulization)*

Dose: composed of value and unit for drug dose *(0.5 mg, four doses, one to two pills, three million units)* or rates *(5 mg/kg)*. If a dose was changed according to a past condition, the modification is annotated among two normalized values (increase, decrease)

2.3.5 Temporal data

Date: absolute and relative dates *(January 2005)*

Moment: moment of a day for drug intake or surgical intervention *(at bedtime, the morning)* or specific time during the hospital stay *(at D1-D2)*

Duration especially for treatments and diseases *(since 10 years, for four weeks)*

Frequency for intakes, diseases, signs and symptoms *(once a day, if needed, chronic, every two weeks)*

2.3.6 Lab and exam results

This category is related to all numerical values from lab results *(105/80 mm Hg, 68 bpm)* and analysis result from examination (e.g., *normal* for imaging or palpation).

2.4 Additional information

Some categories are annotated with additional information.

2.4.1 Linguistic annotations

Similarly to Uzuner et al. (2011), we added assertion values among the six tags possible: *present, absent, associated to someone else, conditional,*

hypothetical, possible. Present: default value; *Absent*: element planned but not realized; *Conditional*: element that can occur under certain circumstances; *Hypothetical*: element that may occur in the future; *Possible*: element that may occur; *Associated to someone else*: element concerning family or acquaintances. Assertions may be used for the annotation of the Pathology, Signs and Symptoms, Tests, and Treatments categories.

2.4.2 Medical information

Linguistic interpretation: With Substances and Weight, if the medication or the weight change according to their previous values, this modification is annotated according to two normalized values: *stop* and *titration* for Substances, and *gain* and *loss* for Weight.

Medical interpretation: For lab results (e.g., blood pressure) and physiological data (temperature), if values can be compared to known ranges (external medical knowledge), three normalized levels are used *(high, normal, low)* in order to provide a better comprehension of those values.

3 Annotated corpus

3.1 Inter-annotator agreement

The inter-annotator agreement is computed with Cohen's κ, and with Precision, Recall and F-measure values (Sebastiani, 2002).

General information We computed inter-annotator agreement scores on the normalized values for general information: Age, Gender and Outcome, and on the annotated text spans for Origin. We achieved excellent agreements for Age and Gender (κ=0.939), differences being due to omissions; poor agreement for Outcome (κ=0.369) due to differences of interpretation between close values (e.g., recovery vs. improvement for long-term diseases); and very low agreement for Origin (κ=-0.762) since spans of text were often distinct between annotators. As stated by Grouin et al. (2011), the κ metric is not well suited for annotations of text since it relies on a random baseline for which the number of units that may be annotated is hard to define. As a consequence, the classical F-measure is often used as an approximation of inter-annotator agreement. In the following experiments, we present the inter-annotator agreements through Precision, Recall, and F-measure.

Outcome The outcome value is complex since differences between recovery and improvement may imply more knowledge than the information presented in the clinical case. As an example, for a patient presenting arterial hypertension at the consultation, do we consider a "recovery" or an "improvement" when clinicians indicate *a complete remission 18 months after the intervention*? Can we consider a recovery for a remission? Is a period of eighteen months sufficient to take a decision? If *no tumor recurrence after fifteen months of decline* is considered, since a tumor may appear again, can we still consider a "recovery"?

At last, we made a difference between cancers or malign tumors ("improvement") and benign tumors or other diseases ("recovery"). For chronic diseases, we only considered an "improvement".

Fine-grained categories In Table 2, we indicate the inter-annotator agreement for the main categories from fine-grained annotations on a subset of 70 clinical cases we annotated in duplicate.

Category	P	R	F
Anatomy	0.5660	0.8511	0.6799
Concentration	0.5714	0.2857	0.3810
Date	0.7042	0.2747	0.3953
Devices	0.3151	0.8519	0.4600
Dose	0.3744	0.8913	0.5273
Duration	0.7500	0.5816	0.6552
Examen.	0.4260	0.8267	0.5623
Function	0.5135	0.2879	0.3689
Frequency	0.5597	0.8824	0.6849
Localisation	0.4328	0.8056	0.5631
Mode	0.5563	0.8778	0.6810
Pathology	0.2596	0.6116	0.3645
SOSY	0.5567	0.6888	0.6157
Specialty	0.3077	0.2051	0.2462
Substance	0.5950	0.7163	0.6500
Treatment	0.5378	0.4054	0.4623
Overall	0.4426	0.6924	0.5400

Table 2: Inter-annotator agreement for the main categories (fine-grained annotations) on the 70 files dataset

We observe that the categories yield better Recall than Precision, which means that similar units are annotated by the two annotators. Yet, Precision values are often lower because the units may correspond to different text spans. The average agreement in terms of F-measure is 0.5400. This first round of fine-grained annotations permitted to elaborate strong annotation guidelines, which

Physiology								Numerical Values
Body measurements			*Vital signs*		*Biology*			
Weight	Size	Surface	Temp.	Liquid	Anatomy	Local.	Function	
8	5	3	5	47	424	603	37	310

Surgery					**Diseases**			
Speciality	Tests	Treatm.	Access	Devices	Pathology	SOSY	Organism	Nature
26	784	251	20	73	285	803	11	192

Drugs					**Temporal**			
Class	Substance	Conc.	Mode	Dose	Date	Moment	Duration	Frequency
44	437	14	142	219	71	174	76	134

Table 3: Number of annotations for each fine-grained category within the subset of 70 files. (Temp.=temperature, Local.=localization, Treatm.=surgical treatments, Access=surgical approach, SOSY=signs or symptoms, Class=pharmaceutical class, Conc.=concentration)

is being applied to the whole set of 717 clinical cases. We expect that the further annotations will provide with better inter-annotator agreement.

3.2 Statistics

Table 3 indicates the number of annotations for each fine-grained category based on a subset of 70 cases. The total number of annotations is 5,198, which gives on average 74.3 annotations per case.

As shown on figure 1, all fine-grained categories have not been used in each file. Six categories are mainly used in the dataset of 70 files: Test (annotations found in 95.7% of all files), Localisation (90.0%), Sign or Symptom (78.6%), Anatomy (75.7%), Pathology (72.9%), and Surgical Treatment (68.6%).

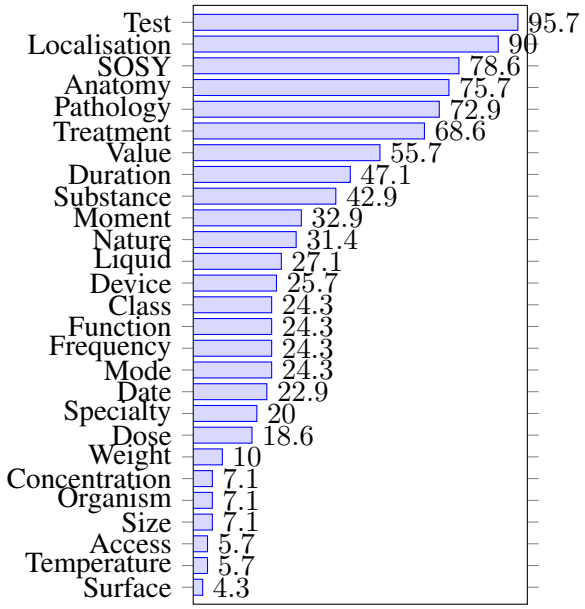

Figure 1: Distribution of fine-grained annotations in the dataset of 70 files (percentage)

Physiological information (body measurements and vital signs) are found in a few number of files (less than 10% of files from the dataset). Since those types of information are useful for a limited number of pathologies or signs or symptoms, they have been found in few documents.

Table 4 presents the final number of annotations on the four general categories and their distribution on the whole dataset of 717 files. Since a few clinical cases describe several patients (either a cohort of patients or a pathology affecting several patients), the total number of annotations may be higher than the total number of files in the corpus. This has been observed for Gender and Origin.

Category	#	Distribution
Age	717	from new born to 98 y.o.
Gender	727	317 feminine, 410 masculine
Outcome	678	227 recovery, 256 improvement, 55 stable, 23 worsening, 117 death
Origin	722	722 distinct spans of text

Table 4: Number of mentions for the general information annotations on the whole dataset of 717 files

Nevertheless, apart from the Gender category, other general annotations are not found in all files: Origin is present in 716 files (99.9% of files), Age in 698 files (97.4%), and Outcome in 675 files (94.1%). Annotations are missing when it was not possible to identify the information.

3.3 Annotated clinical case report

Figure 2 shows the following clinical case: *A 73-year-old woman who had only one child by caesarean section, but had for several years a*

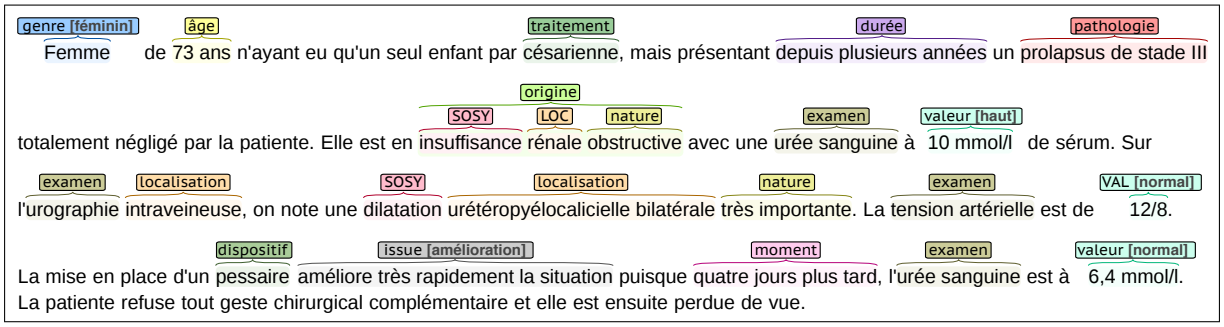

Figure 2: Annotated case report. General information includes the following tags: *genre* (gender), *âge* (age), *origine* (origin), *issue* (outcome). Other tags are related to fine-grained information. Normalized values appear between square brackets (feminine gender, high or normal values, improvement outcome)

stage III prolapse totally neglected by the patient. She is in obstructive renal failure with blood urea at 10 mmol/l serum. On the intravenous urography, we notice a very significant bilateral ureteropyelocaliceal dilation. The blood pressure is 12/8. The pessary placement very quickly improves the situation since four days later, the blood urea is 6.4 mmol/l. The patient refuses any additional surgery and is then lost to follow-up. The case is annotated with general and fine-grained information. Elements in square brackets correspond to normalized tags: feminine *("féminin")* for gender, high *("haut")* and normal for values, and improvement *("amélioration")* for outcome.

3.3.1 Types of information in the typical clinical case report

This case report is composed of several parts, annotated as follows:

- *general description with patient history:* gender (woman, *"femme"*); age (73-year-old, *"73 ans"*); surgical treatment (caesarean, *"césarienne"*); duration (for several years, *"depuis plusieurs années"*); pathology (stage III prolapse, *"prolapsus de stade III"*)

- *origin of consultation, tests and results:* origin (obstructive renal failure, *"insuffisance rénale obstructive"*), composed of three elements: sign or symptom (failure, *"insuffisance"*), localization (renal, *"rénale"*), and nature (obstructive, *"obstructive"*); three tests (blood urea, *"urée sanguine"*, blood pressure, *"tension artérielle"*, urography, *"urographie"*) with lab results (10 mmol/l, 12/8, 6.4 mmol/l) and localization (intravenous, *"intraveineuse"*); sign or symptom (dilation, *"dilatation"*) with localization (bilateral ureteropyelocaliceal, *"urétéropyélocalicielle bilatérale"*) and nature (very significant, *"très importante"*)

- *surgical treatment and issue:* medical device (pessary, *"pessaire"*); outcome (very quickly improves the situation, *"améliore très rapidement la situation"*); moment (four days later, *"quatre jours plus tard"*)

- *follow-up:* no annotation in this clinical case

We observe the types of information contained in clinical case reports are similar to those typically provided by patient health documents in hospitals.

3.3.2 Distribution of annotations

Columns two and three from Table 4 indicate that general information are found in all clinical cases. For gender and origin, the number of annotations if higher than the number of clinical cases because several people are described in some cases (gender), and because several origins of consultation may be indicated (namely, several signs or symptoms).

From Table 3, one can observe a very imbalanced number of annotations per category. The main categories are: signs or symptoms (15.4%), tests (15.1%), localizations (11.6%), substances (8.4%), and anatomical parts (8.2%). The number of signs and symptoms mentions are three times higher than annotations of diseases (5.5%). Small categories are related to specific data (especially body measurements and vital signs) that are indicated in a limited number of cases. This may correspond to the average difference with the clinical patient reports.

4 Experiments and analysis

The annotated corpus has been exploited to perform similar annotations automatically and for their evaluation. Our aim is to verify the adequateness of the annotations for this information extraction task, as well as to serve as baseline for future work. We specify we do not aim to provide new methods, nor to improve existing systems, but to present a few use cases that may be done on the annotations presented in section 2.

4.1 Linguistic analysis

Syntax. Depending on the outcome observed in clinical cases, we studied the distribution of a few verbal tenses based on the POS annotations provided by the TreeTagger system (Schmid, 1994). As presented in Table 5, past perfect is the main tense for death outcome while present is the main tense for both improvement and stable condition outcomes. Conversely, we observe no future tense in case reports concerned by death.

Verbal tense	R	I	S	W	D
future	0.01	0.01	0.01	**0.02**	0.00
imperfect	**0.19**	0.15	0.12	0.13	0.16
past perfect	0.41	0.41	0.41	0.41	**0.45**
present	0.23	**0.29**	**0.29**	0.24	0.26

Table 5: Percentage of verbal tenses use depending on the outcome value (R=recovery, I=improvement, S=stable, W=worsening, D=death)

Table 6 presents the distribution of demonstrative pronouns (PRO:dem) vs. personal pronouns (PRO:per) depending on the outcome. We observe that impersonal linguistic constructions are mainly used for stable condition outcomes (less personal pronouns and more demonstrative pronouns) than in other outcome types, as if the uncertainty of the stable condition (no improvement nor worsening) would prevent from a too much personal representation of the case.

POS tag	R	I	S	W	D
PRO:dem	0.19	0.19	**0.24**	0.21	0.18
PRO:per	**0.52**	0.50	0.47	0.48	0.51

Table 6: Percentage of types of pronoun use (PRO:dem=demonstrative, PRO:per=personal) depending on the outcome value (R=recovery, I=improvement, S=stable, W=worsening, D=death)

Semantics. Table 7 presents the main elements annotated as anatomical parts, pathologies, signs or symptoms, and surgical treatments depending on the gender. The observed differences of medical entities mainly highlight differences due to anatomical parts specific to men or women, or to distinct prevalences of pathologies. We observe less differences in surgical treatments than in other categories.

Category	F/M	Annotated spans
Anatomy	F	kidney, bladder, torso
	M	testicle, bladder, prostate
Pathology	F	acute pyelonephristis, adenocarcinoma, carcinoma, edema, mydriasis, tumor
	M	adenocarcinoma, fistula, rhabdomyosarcoma, tuberculosis, tumor, ulcer
Signs or Symptoms	F	dilation, hematuria, hypersensitivity, lesion, mass, pain, rash, stone, vomiting
	M	fever, infection, lesion, mass, nodule, pain, pneumonia, relapse, retention, trouble
Treatments	F	chemotherapy, curettage, desensitization, exeresis, lumpectomy, nephrectomy
	M	ablation, chemotherapy, clamping, desensitization, exeresis, orchiectomy, plasma exchange, resection

Table 7: Most used anatomical parts, pathologies, signs or symptoms, and surgical treatments depending on the gender (F=feminine, M=masculine)

4.2 Information extraction

The information extraction experiments rely on the Wapiti tool (Lavergne et al., 2010) that implements linear chain CRF (Lafferty et al., 2001). We trained a model on the 16 fine-grained categories presented in Table 2, through a 10 fold cross-validation process, using a $l1$ regularization. We used the following features: unigrams and bigrams of tokens, number of characters, typographic case, presence of punctuation and digit, Soundex code[1] value of each token, relative position of token within the document (beginning, middle, end), POS tags from the TreeTagger sys-

[1] https://en.wikipedia.org/wiki/Soundex

tem (Schmid, 1994) and syntactic chunks based on those tags, presence of the token in a dictionary of 251k inflected forms for French, and cluster id (120 classes) of each token using the clustering algorithm from Brown et al. (1992) implemented by Liang (2005). The results that we achieved are presented in table 8. Overall, we obtain 0.76 Precision, 0.45 Recall and 0.67 F-measure.

Category	P	R	F
Anatomy	0.7260	0.4823	0.5795
Concentration	0.5000	0.0714	0.1250
Date	1.0000	0.4507	0.6214
Devices	0.3077	0.0548	0.0930
Dose	0.7805	0.5818	0.6667
Duration	0.9545	0.2692	0.4200
Examen.	0.8308	0.6303	0.7168
Function	0.8889	0.2162	0.3478
Frequency	0.9630	0.1955	0.3250
Localisation	0.7812	0.5795	0.6654
Mode	0.8929	0.5245	0.6608
Pathology	0.5918	0.2086	0.3085
SOSY	0.6067	0.3639	0.4549
Specialty	1.0000	0.3846	0.5556
Substance	0.8490	0.3721	0.5175
Treatment	0.8190	0.3785	0.5177
Overall	0.7640	0.4492	0.5658

Table 8: Results achieved using a CRF through a 10 folds cross-validation

5 Discussion

Corpus. One contribution of this work is related to the availability of the annotated corpus from the medical domain for French. We based our annotation schema on both existing ones (semantic types from the UMLS, i2b2 NLP Challenges) and on types of elements found in our corpus. This annotated corpus will be made available for the research purposes and may be of interest for several NLP tasks related to the biomedical domain: information extraction, relationships identification, classification, discourse analysis, temporality, etc.

Human annotations vs. CRF. We observed that results obtained by the designed CRF system are in line with results obtained by humans when annotating the corpus. More specifically, while humans were producing the gold standard, they had to deal with categories harder to process than others. We also observe that those categories are generally difficult to retrieve and annotate with the CRF model as well: Concentration (F=0.38 vs. 0.13), Function (F=0.37 vs. 0.35), and Pathology (F=0.36 vs. 0.31). An explanation is the lack of regularity (for the CRF system) and ambiguous content w.r.t. content from other categories.

Yet, two categories considered as hard for humans yielded better results than expected with the CRF model: Specialty (F=0.25 vs. 0.56) and Dates (F=0.40 vs. 0.67). The differences observed between humans which produce those bad results were mainly due to omissions. Conversely, humans outperformed the CRF model on Frequency (F=0.68 vs. 0.33), Duration (F=0.66 vs. 0.42), and Devices (F=0.46 vs. 0.09). Those categories are composed of distinct elements with low frequencies of use which are complex to process for a probability-based system, but basic for humans.

As future work, we plan to continue the fine-grained annotation of the whole corpus. We also plan to define relationships between the existing entities, in order to provide annotations of relations. Despite the absence of relationships annotations, the corpus can still serve to perform unsupervised experiments. Such results may be used for automatic pre-annotation of relationships, in order to make it easier the human annotation work.

6 Conclusion

In this paper, we presented a corpus composed of 717 medical clinical case reports, written in French, with two levels of annotations (general and fine-grained annotations). Our annotation schema is composed of four general categories (age, gender, outcome, origin) for a total of 2,835 annotations, and 27 fine-grained categories dealing with five domains (physiology, surgery, diseases, drugs, temporal) for a total of 5,198 annotations on a subset of 70 files. For certain categories, the annotations are provided under a normalized format (age, gender, outcome) while other categories are associated with additional information based on a human judgement, either of linguistic nature (assertions, change of conditions) or medical nature (lab results compared to known ranges). The corpus and its annotations will be made available for the research. We expect that the availability of this corpus may boost the research on biomedical textual data in French, and provide the domain with more robust and stable tools leading to a better reproducibility of the results.

Acknowledgments

This work has been founded by the French ANR (grant number ANR-17-CE19-0016-01) as part of the project CLEAR (Communication, Literacy, Education, Accessibility, Readibility) and by the French government support, granted to the CominLabs LabEx, managed by the ANR in Investing for the Future program (grant number ANR-10-LABX-07-01).

References

Dwight Atkinson. 1992. The evolution of medical research writing from 1735 to 1985: The case of the Edinburgh medical journal. *Applied Linguistics*, 13(4):337–74.

Peter F Brown, Vincent J Della Pietra, Peter V de Souza, Jenifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonardo W D'Avolio, Guergana K Savova, and Özlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, 18(5):540–3.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc*, 24(3):596–606.

David Doermann and Shee Yao. 1995. Generating synthetic data for text analysis systems. In *Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. ArXiv:1903.11508v1.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. Cas: French corpus with clinical cases. In *Proc of LOUHI*, pages 122–128, Brussels, Belgium.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc of Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen-Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform*, 45(5):885–92.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Mehmet Kayaalp. 2017. Modes of de-identification. In *AMIA Annu Symp Proc*, pages 1044–50, San Francisco, USA.

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

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McRay. 1993. The Unified Medical Language System. *Methods Inf Med*, 32(4):281–91.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution – A case study to get around IPRs and privacy constraints featuring the German JSynCC corpus. In *Proc of LREC*, pages 1259–66, Miyazaki, Japan.

Yuliia Lysanets, Halyna Morokhovets, and Olena Bieliaieva. 2017. Stylistic features of case reports as a genre of medical discourse. *J Med Case Rep*, 11:83.

Stéphane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Method*, 10(70).

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural languague processing: an introduction. *J Am Med Inform Assoc*, 18(5):544–51.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Richard A Rison. 2013. A guide to writing case reports for the journal of medical case reports and biomed central research notes. *J Med Case Rep*, 7:239.

Richard A Rison, Jennifer Kelly Shepphird, and Michael R Kidd. 2017. How to choose the best journal for your case report. *J Med Case Rep*, 11:198.

Yoichi Satomura and Marcio Biczyk Do Amaral. 1992. Automated diagnostic indexing by natural language processing. *Med Inform*, 17(3):149–63.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.

F Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *J Am Med Inform Assoc*, 17(5):514–8.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–6.