# Relation Extraction from Clinical Cases for a Knowledge Graph

Agata Savary, Alena Silvanovich, Anne-Lyse Minard, Nicolas Hiot, Mirian Halfeld Ferrari

# Relation Extraction from Clinical Cases for a Knowledge Graph

Agata Savary[1][0000−0002−6473−6477], Alena Silvanovich[2], Anne-Lyse Minard[3][0000−0001−6197−0463], Nicolas Hiot[2,4][0000−0003−4318−4906], and Mirian Halfeld-Ferrari[2][0000−0003−2601−3224]

[1] LISN, Paris-Saclay University, France
[2] LIFO, Université d'Orléans, INSA CVL, France
[3] LLL, Université d'Orléans, France
[4] EnnovLabs – Ennov, Paris, France
agata.savary@universite-paris-saclay.fr, alentelvy@gmail.com, {anne-lyse.minard, mirian}@univ-orleans.fr, nhiot@ennov.com

**Abstract.** We describe a system for automatic extraction of semantic relations between entities in a medical corpus of clinical cases. It builds upon a previously developed module for entity extraction and upon a morphosyntactic parser. It uses experimentally designed rules based on syntactic dependencies and trigger words, as well as on sequencing and nesting of entities of particular types. The results obtained on a small corpus are promising. Our larger perspective is transforming information extracted from medical texts into knowledge graphs.

## 1 Introduction

Transforming data into information and then into knowledge is the focus of our action. In there, one of the main challenges concerns the construction of a graph database instance from a set of textual documents, sometimes referred to as the construction of a *knowledge graph*. This paper deals with a step towards this goal.

Nowadays, knowledge graphs are considered essential to allow smart data exploitation. They are intended to use an *organizing principle* so that a user (or a computer system) can reason about the underlying data. This organization principle puts in place a meta-data level (a schema) that adds context to knowledge discovery.

Our goal is to work on attributed graph databases which become very popular both in industry and academia [9]. This comprises nodes, representing entities (such as people, drugs, and exams), and edges, representing relationships between the entities. Graph databases are to be used when the relationships are as important as the entities themselves. Any number of attributes (also called properties), in the form of key-value pairs, may be associated with the nodes or the edges.

The ultimate goal of our work is to automatically map text to a given schema, building in this way a database instance that respects that schema. The schema here can be built as a collaborative task where techniques from natural language and database model design interact. Our corpus is a collection of clinical cases from which we would like to extract entities (classes) and relationships among them. The following example illustrates our general propose.

*Example 1.* Let us consider the following clinical case extract[5]: "*A female patient in the age group 55–60 years presented to us with blurring of vision in both eyes. On slit-lamp examination, numerous circular to oval fleck-like discrete blue opacities at the level of deep corneal stroma and Descemet's membrane was observed.*" In this example, we want a database instance to represent patients having some symptoms and examinations they pass. Let us consider the representation of Figure 1 where *Patient*, *Anatomy*, *Symptom* and *Examination* are types of nodes and edges represent relationships *PassExam*, *HasSymp*, *GivesRes* and *ConcernsAnat*. To structure the information, we can consider a schema designed by a database designer from the information obtained through natural language processing. This conception is a big challenge. One should consider many different aspects of the problem, starting with questioning what information is really important to store. The solution also depends on the (analytical) queries we want to consider later on. Then, for a graph database, we should also decide what information is represented as a node, a property or a relationship. Usually entities give rise to nodes, but the distinction between properties and relationships is not evident. Choices may impact the efficiency of query processing. □

It should be clear that we do not want just to transform textual structures into a graph – some tools exist to represent a text corpus as a graph[6] [10,17], but they do not go further, trying to build a higher-level model. Here, the idea is to add or infer metadata (the schema) and organize the information originally available in texts according to this higher abstraction. Different proposals exist as, for instance, to use generic taxonomies and ontologies as the property graph model. Our work focuses on custom data models: a public standard can be used as a starting point, but we let the database designer introduce her own organisation principles. In this scenario, the information extraction pipeline is usually composed of the following steps: named entity recognition and classification, co-reference resolution, and relationship extraction. These steps should deal with well known challenges in Natural Language Processing (NLP), such as disambiguation or temporal representations.

In this paper, we focus on relation extraction in NLP to contribute to the construction of a knowledge graph. In our previous work, we have concentrated our attention on entities: in [19] we propose a method for the extraction of nested entities that uses a cascade of Conditional Random Fields (CRF); [7] in [2] we propose entity enrichment in order to translate natural language queries into database queries. In fact, from the enrichment of an entity some relationships can be detected. In the current paper, we consider the extraction of entities as done in [19], and *we investigate the extraction of binary relations between them, broadening the initial ideas we used in [2].*

It is worth noticing that the gap between the mentioned NLP steps and the wanted data model is still big. The construction of such a model needs 'external' semantic information (standard or customized) and should always be driven by its intended use. The final efficiency of the model is essential.

[5] Extract from PubMed: `https://pubmed.ncbi.nlm.nih.gov/35365471/`

[6] `https://www.slideshare.net/lyonwj/natural-language-processing-with-graph-databases-and-neo4j`

[7] Conditional Random Fields [15] are probabilistic models often used in NLP for sequence labelling tasks as they take into account the context of the samples to label.

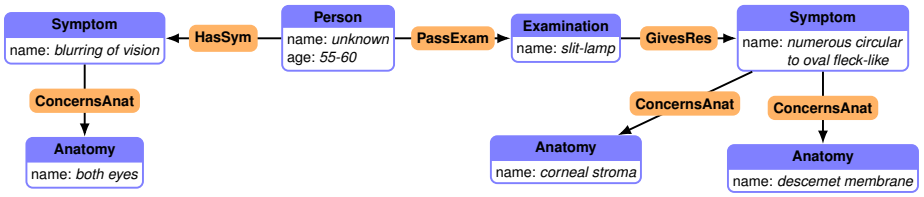


Fig. 1: Example of a Neo4J graph database instance we would like to obtain from the given text. Most properties of edges and nodes are not shown.

*Paper Organization.* Sec. 2 presents related works. Sec. 3 describes the medical corpus used in the experiments and summarizes the entity extraction method developed in previous work. Sec. 4 addresses the typology of relations between entities and Sec. 5 presents the rule-based method for extracting these relations from the corpus. Experimental results are shown in Sec. 6 and Sec. 7 closes the paper.

## 2 Related Work

Relation extraction is a task of information extraction which consists in detecting if two or more entities are linked and in classifying the link. In this work we are interested only in binary relations, i.e. relations between two entities, and we want both to detect and to classify them.

The first methods developed for relation extraction where based on patterns, manually defined or automatically extracted [14]. In the medical domain we can cite the SemRep system [22] which used the UMLS in order to define semantic patterns, [1] used lexicalized patterns and [7] resorted to multi-layer patterns (word forms, lemmas and parts-of-speech). The RelEx system [11], designed to extract protein-protein interaction relations, made use of rules based on dependency trees. First, the dependency path that linked two entities of interest (proteins) was extracted, then the dedicated rules identified relations of three types: effector-relation-effectee, relation-of-effectee-by-effector, relation-between-effector-and-effectee.

Supervised machine learning methods are used when the size of the annotated training data is big enough. In the biomedical domain, the most popular techniques are SVM (e.g. [23] [18]) and CRF, with a large variety of features (surface, semantic, syntactic, etc.). [6] hypothesized that "instances containing similar relations will share similar substructures in their dependency trees". Therefore they developed a system based on SVM and augmented dependency trees (addition of features on the nodes, such as part-of-speech, chunk tag, etc.) in order to extract relations from newswire documents (located, citizen-of, part-of, etc.). Their experiments show an improvement of performances using dependency tree kernels instead of bag-of-words kernels. More recently, a lot of methods have been proposed based on deep learning techniques (for example [16]). As in our task we do not have any training data, we cannot use these supervised methods.

In order to counter the issue of the availability of annotated data, we had a look at unsupervised methods. The Open Information Extraction task enables to extract rela-

tions from frequent syntactic patterns, in particular using the subject-verb-object structure (see ReVerb [8]). It assumes that the relations of interest are always expressed by similar simple syntactic structures or by verbs. We have observed complex syntactic structures in our corpus, and openIE systems did not bring usable results. Finally, [20] take advantage of an existing database in order to automatically build an annotated corpus from a projection of the relations saved in it. This method is called distance supervision and requires a database instead of a manually annotated corpus. In our project, we do not have a database on input, and the goal is precisely to automatically build one, so we cannot use distance supervision.

## 3 Corpus and entity extraction

This work is based on the CAS corpus - French Corpus with Clinical Cases [12] - used in multiple editions of DEFT (*Défi Fouille de Textes*) [5,13], an evaluation campaign of systems dedicated to French medical text processing. The corpus is composed of 167 clinical cases in French (100 cases for training and 67 for testing) and it contains 13 548 entities. The clinical cases came from publications in scientific literature and teaching samples for medical students. The cases are anonymized and describe real or fictitious situations from different medical domains (cardiology, urology, oncology, etc.). Some parts of the corpus were manually annotated from scratch (by two annotators), while others were automatically pre-annotated (by CRF-based methods) and then manually corrected.

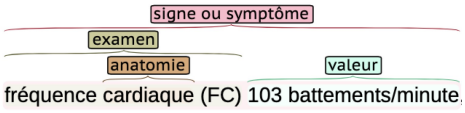


Fig. 2: Nested entities in the CAS corpus.Trans.: *'heart rate (HR) 103 beats/minute'*

The corpus contains annotation for 13 types of entities: 2 for temporal expressions (DATE, MOMENT), which will be ignored in this paper, 4 for medical objects and facts (ANATOMIE 'ANATOMY', EXAMEN 'EXAM', PATHOLOGIE 'PATHOLOGY', and SIGN OU SYMPTÔME - SOSY 'SIGN OR SYMPTOM - SOSY'), and 7 for patient treatment (DOSE, DURÉE 'DURATION', FRÉQUENCE 'FREQUENCY', MODE, SUBSTANCE, TRAITEMENT 'TREATMENT', and VALEUR 'VALUE').[8] Shorter entities can be nested in larger ones, for example an exam entity can contain the body part (anatomy) on which the exam is carried out (Fig. 2). The statistics of the entity types and their nesting are shown in Tab. 1 for the subset of CAS corpus used in this paper.

In previous work [19], we used the CAS corpus for automatic entity extraction with a CRF cascade. This approach was motivated by the frequent nesting of entities (e.g. 1831 SOSY entities contain 1909 nested entities). The idea was to extract entities of each

[8] Entity types are listed in French and with the English translation, if it differs from French. In the rest of the paper, the English names are used in the core of the texts, while the French ones appear in figures. The entity types SUBSTANCE, ANATOMY and TREATMENT will sometimes be abbreviated by SUB, ANAT and TREAT, respectively.

nesting level with a different CRF model, so that the output of earlier CRF layers is used on input of further layers. The method showed an overall precision of $0.839$, a recall of $0.613$ and an F1 score of $0.708$. These performances allow us to expect reasonable quality of automatic entity extraction in new in-domain texts.

The situation is, however, more difficult as far as relation extraction is concerned, since the CAS corpus contains no annotations for relations. Therefore, we constructed our own small development corpus. A typology of relations to annotate was first defined (cf. Section 4). Then, the 11 longest documents in CAS were selected and annotated by 4 annotators (each text was manually annotated by a single annotator, and some texts were double-checked by another one). The resulting corpus is composed of $6289$ words, and contains annotations for $1421$ entities and $742$ relations. It was split into a development and a test sub-corpus of respectively $6$ and $5$ files.

| Entity | Number | Nested entities |
|---|---|---|
| SOSY | 277 | SUB, ANAT, EXAM, VALUE |
| SUBSTANCE | 26 | |
| ANATOMY | 212 | |
| EXAM | 146 | SUB, ANATOMY |
| DOSE | 125 | |
| MODE | 96 | |
| FREQUENCY | 88 | |
| VALUE | 85 | |
| PATHOLOGY | 55 | ANAT, VALUE |
| TREATMENT | 44 | ANAT |
| DURATION | 25 | |
| Total | 1421 | |

Table 1: Entities in the subset of the CAS corpus used in this paper

Finally, since our relation extraction methods rely notably on syntactic patterns (cf. Section 5), the CAS corpus was pre-processed with the SpaCy[9] parser using the `fr_core_news_md` model[10] trained on the UD French Sequoia v2.8 corpus[11]. This corpus contains, in particular, a number of texts from the European Medicines Agency.

## 4 Typology of relations

To define a typology of relations, we started from existing annotation schemas for medical texts in French, in particular [4] which is mainly based on UMLS [3] and composed of 37 relations classified into 5 types: aspect relations, assertion relations, drug-attribute relations, temporal relations and event-related relations, and [21] which uses 10 relations, defined jointly with experts in radiology: localisation, target, sign, cause...

Among these relations we selected those which: (i) are binary, (ii) occur frequently in texts, (iii) are specific enough to apply to a low number of entity types, (iv) have stable behavior (to be able to extract them with rules). We decided not to work on temporal and causal relations for the moment. They are generic (not domain-specific), more complex to extract and covered by a separate task in information extraction. The resulting set contains 5 relations: MOYEN 'MEANS', MESURE_DE 'MEASURE_OF', ACCOMPAGNE 'ACCOMPANIES', RÉLÈVE/RECHERCHE/TESTE 'REVEALS/SEARCHES/TESTS' and LOCALISATION 'LOCATION'. Tab. 2 shows the 5 relations as well as types of entities to which they apply.

[9] `https://spacy.io/`
[10] `https://spacy.io/models/fr`
[11] `https://github.com/UniversalDependencies/UD_French-Sequoia`

| Relation's name | Entity type pairs | Examples |
|---|---|---|
| MEANS | MODE-SUBSTANCE | *une* [*crème*]$^{\text{MODE}}$ *de* [*nistatin*]$^{\text{SUB}}$ *a été préscrite*<br>'an [ointment]$^{\text{MODE}}$ of [nistatin]$^{\text{SUB}}$ was prescribed' |
| MEASURE_OF | DOSE-SUBSTANCE<br>VALUE-SUBSTANCE | *la* [*doxorubicine*]$^{\text{SUB}}$ *à raison de* [*37,5 mg/m2/dose*]$^{\text{DOSE}}$<br>'[*37,5 mg/m2/dose*]$^{\text{DOSE}}$ of [doxorubicine]$^{\text{SUB}}$, |
| ACCOMPAGNIES | PATHOLOGY-PATHOLOGY<br>SOSY-SOSY<br>PATHOLOGY-SOSY<br>SUBSTANCE-SUBSTANCE | *le patient présentait une* [*fatigue importante*]$^{\text{SOSY}}$ *de même qu'une* [*hyperthermie*]$^{\text{SOSY}}$<br>'the patient displayed [great tiredness]$^{\text{SOSY}}$ as well as [hyperthermia]$^{\text{SOSY}}$, |
| REVEALS/ SEARCHES/ TESTS | EXAM-VALUE<br>EXAM-SOSY<br>EXAM-PATHOLOGY | *l'*[*échographie abdominale et pelvienne*]$^{\text{EXAM}}$ *révèle la présence d'une* [*masse surrénalienne à droite*]$^{\text{SOSY}}$<br>'the [abdominal and pelvic ultrasonography]$^{\text{EXAM}}$ reveals a [growth in the right adrenal gland]$^{\text{SOSY}}$, |
| LOCATION | ANATOMY-PATHOLOGY<br>ANATOMY-SOSY<br>ANATOMY-EXAM<br>ANATOMY-TREATMENT | *elle a subi une* [*résection au niveau du* [*lobe supérieur droit*]$^{\text{ANAT}}$]$^{\text{TREAT}}$<br>'she underwent a [resection of the [upper right lobe]$^{\text{ANAT}}$]$^{\text{TREAT}}$, |

Table 2: Relations treated in this work and types of entities to which they apply.

## 5 Rule-based relation extraction

The aim of this work is to automatically extract relations between entities in a medical text, to feed a knowledge graph. Most of the state-of-the-art methods in relation extraction (Sec. 2) are hardly applicable in our case due to the lack of data. Medical texts are most often concerned by severe privacy constraints and are rarely available outside of the strictly clinical context. The CAS corpus, the one we use, is one of the rare (if not the only) French dataset of clinical cases available for NLP research. It is not annotated for relations, thus, supervised relation extraction methods are currently excluded. Non-supervised methods, such as OpenIE, are not appropriate either, since they do not apply to a pre-defined set of relations and have low precision. Finally, distant supervision requires a large pre-existing knowledge base of relevant relations, which is not available for the medical domain in French. Under these strong data constraints, we resorted to rule-based methods relying on syntactic, lexical and surface clues.

***Syntactic rules*** The hypothesis behind the syntactic rules is that syntactic dependencies between words signal semantic relations between entities containing these words. For instance, Fig. 3 shows a one-word entity [*dyphenhydramine*]$^{\text{SUB}}$ and a 2-word entity [*voie IV*]$^{\text{MODE}}$ '[intraveneous route]$^{\text{MODE}}$' connected by the MOYEN 'MEANS' relation. The corresponding dependency graph reveals that a syntactic dependency exists between the headwords (*dyphenhydramine* and *voie* 'route') of these two entities.

Note that (in Tab. 2) the type of the relation between two entities (if any) is fully determined by the types of these entities. Entities of types MODE and SUBSTANCE can only occur in a relation of type MEANS, those of types DOSE and SUBSTANCE in a relation MEASURE_OF, etc. Given two entities $E_1$ and $E_2$ of types $T_1$ and $T_2$, respectively, as well as a relation type $R$, we will say that $E_1$ and $E_2$ are $R$-compatible if $R$ can occur between entities of types $T_1$ and $T_2$ according to Tab. 2. For instance,

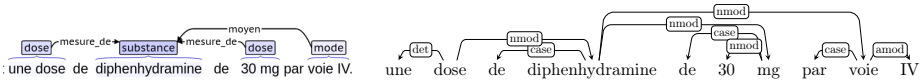


Fig. 3: The syntactic dependency of type NMOD (nominal modifier) between *diphenhydramine* and *voie* 'route' signals the semantic relation of type MEANS between a mode and a substance. Trans.: *a dose of dyphenhydramine of 30 mg via intravenous route*


Fig. 4: The common head *reçoit* 'receives' signals the relation of type MEASURE_OF between a dose and a substance. Also, the occurrence of *300 mg* between *acétaminophène* 'acetaminophen' and *voie orale* 'oral route', as well as a dependency between *mg* and *voie* 'route', signal the relation of type MEANS between a mode and a substance. Trans.: *receives acetaminophen 300 mg by oral route*.

[*dyphenhydramine*]$^{\text{SUB}}$ and [*30 mg*]$^{\text{DOSE}}$ in Fig. 3 are MEASURE_OF-compatible. The above observations and definitions allow us to formulate three generic syntactic rules :

**Syn1** If two entities $E_1$ and $E_2$ are $R$-compatible and a direct syntactic link occurs between any two of their components, then a relation of type $R$ should be inserted between $E_1$ and $E_2$. This rule is illustrated by Fig. 3, with $E_1$, $E_2$ and $R$ equal to [*dyphenhydramine*]$^{\text{SUB}}$ and [*voie IV*]$^{\text{MODE}}$ '[*intraveneous route*]$^{\text{MODE}}$' and MOYEN 'MEANS', respectively.

**Syn2** If two entities $E_1$ and $E_2$ are $R$-compatible and (any two of their components) have the same head[12], then a relation of type $R$ should be inserted between $E_1$ and $E_2$. For instance, in Figure 4, the entities [*acétaminophène*]$^{\text{SUB}}$ and [*300 mg*]$^{\text{DOSE}}$ are MEASURE_OF-compatible. They have incoming dependencies OBL:ARG (oblique nominal) and OBJ (object) outgoing from the same word *reçoit* 'receives', and they are, indeed, connected by a relation of type MEASURE_OF.

**Syn3** If two entities $E_1$ and $E_2$ are $R$-compatible, if a third entity $E_3$ occurs between $E_1$ and $E_2$, and if (any component of) either $E_1$ or $E_2$ has a direct syntactic link with (any component of) $E_3$, then a relation of type $R$ should be inserted between $E_1$ and $E_2$. For instance, in Figure 4, the entities [*acétaminophène*]$^{\text{SUB}}$ '[acetaminophen]$^{\text{SUB}}$' and [*voie orale*]$^{\text{MODE}}$ '[oral route]$^{\text{MODE}}$' are separated by a third entity [*300 mg*]$^{\text{DOSE}}$. A direct dependency (of type NMOD) connects *mg* with *voie* 'route', and there is, indeed, a relation (of type MOYEN 'MEANS') between [*acétaminophène*]$^{\text{SUB}}$ and [*voie orale*]$^{\text{MODE}}$.

While the syntactic rules have a relatively good coverage, they suffer from at least two weaknesses. Firstly, the parsing results may be erroneous: some dependencies may

[12] Word $w_1$ is a syntactic head of word $w_2$ if there is a syntactic dependency link outgoing from $w_1$ and incoming in $w_2$. Most dependency parsing models ensure that each word (except the root of the sentence) has exactly one head, i.e. the dependency graph is a tree.

be missing or spurious. Secondly, the large variety of possible syntactic structures in which entities occur would require a large number of specific rules whose precision might be low. So, we also resorted to lexical rules which abstract away from syntax and look at the relative position and the context in which two entities occur in a sentence.

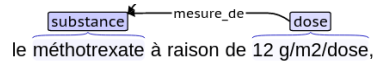


Fig. 5: Relation of type MEASURE_OF between a substance and a dose signaled by the trigger *à raison de* 'at the rate of'. Trans.: *methotrexate at the rate of 12 g/m2/dose*.

***Lexical rules*** A precise relation of type $R$ is sometimes signaled by precise trigger words which occur between two entities. We call such words *$R$-compatible triggers*. For instance, in Figure 5 the relation of type MEASURE_OF between the non-adjacent entities [*méthotrexate*]$^{\text{SUB}}$ and [*12 g/m2/dose*]$^{\text{DOSE}}$ is signaled by the trigger sequence *à raison de* 'at the rate of'. We experimentally established short lists of triggers for two relations: MEASURE_OF and ACCOMPANIES. They are listed in Tab. 3. This, allows us to formulate the following lexical extraction rule:

**Lex1** If entities $E_1$ and $E_2$ are $R$-compatible and an $R$-compatible trigger occurs between them, then a relation of type $R$ should be inserted between $E_1$ and $E_2$.

| Relation | Trigger words | Example |
|---|---|---|
| MEASURE_OF | *à raison de* 'at the rate of', *dosé* 'measured out', *concentration* | [*cidofovir*]$^{\text{SUB}}$ [*intraveneux*]$^{\text{MODE}}$ ***à raison de*** [*375 mg*]$^{\text{DOSE}}$ '[intraveneous route]$^{\text{MODE}}$ [cidofovir]$^{\text{SUB}}$ **at the rate of** [375 mg]$^{\text{DOSE}}$' |
| ACCOMPANIES | *ainsi que* 'as well as', *sans* 'without' *associé à* 'associated with', *avec* 'with' | *un traitement intraveineux de* [*métoclopramide*]$^{\text{SUB}}$ ***associé à*** de la [*diphénhydramine*]$^{\text{SUB}}$ 'an intraveneous treatment of [metoclopramide]$^{\text{SUB}}$ **associated with** [diphenhydramine]$^{\text{SEB}}$' |

Table 3: Relations detectable by trigger words

The observation of the entities and relations in the CAS corpus led us to formulate and implement other lexical rules, which we finally did not retain due to their weak performances. In particular, we extracted a list of predicates such as *révèle* 'reveals', *confirme* 'confirms', *demeure* 'remains', etc. signaling the REVEALS/SEARCHES/TESTS relation as shown in the examples below. Notice that the variety of such predicates is huge and hinders the reliability of the corresponding lexical patterns.

- L'[*échographie abdominale*]$^{\text{EXAMEN}}$ ne **révèle** [*aucune anomalie*]$^{\text{SOSY*}}$[13]
  The [abdominal ultrasonogram]$^{\text{DOSE}}$ reveals [no anomaly]$^{\text{SOSY*}}$
- l'[*imagerie par résonance magnétique abdominale*]$^{\text{EXAMEN}}$ **confirme** la présence d'une [*masse surrénalienne à droite*]$^{\text{SOSY}}$
  the [abdominal magnetic resonance imaging]$^{\text{EXAM}}$ confirms the presence of a [growth in the right adrenal gland]$^{\text{SOSY}}$'

[13] An asterisk following an entity type signals a negated entity occurrence.

***Sequence rules*** Sometimes a relation $R$ can be detected by the sheer proximity of entities of the relevant types. For instance, Figure 6 shows two entities [*échographie de l'appareil urinaire*]$^{\text{EXAMEN}}$ '[ultrasonogram of the urinary system]$^{\text{EXAM}}$' and [*reins de taille normale*]$^{\text{SOSY}}$ '[kidneys of regular size]$^{\text{SOSY}}$'. Their sequence signals the relation of type REVEALS/SEARCHES/TESTS. We experimentally checked that this relation type can be reliably detected by this principle.

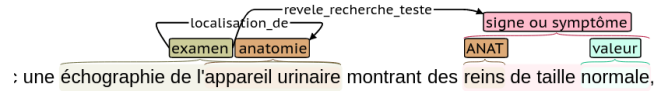


Fig. 6: Entities of type EXAM and SOSY occuring in a sequence and connected by a relation of type REVEALS/SEARCHES/TESTS. Trans.: *ultrasonogram of the urinary system showing the kidneys of regular size*.

A sequence of two entities, connected by a particular relation, can also include a third entity of another specific type. For instance, in Figure 4, the entities [*acétaminophène*]$^{\text{SUB}}$ '[acetaminophen]$^{\text{SUB}}$' and [*voie orale*]$^{\text{MODE}}$ '[oral route]$^{\text{MODE}}$' are separated by entity [*300 mg*]$^{\text{DOSE}}$, and a relation of type MEANS occurs between the first two. While rule Syn3 allows us to extract the MEANS relation in this examples, it misses other cases with an erroneous syntactic analysis. We experimentally checked that the principle of spotting sequences of 3 entities of specific types is especially reliable for the MEANS relation type. Thus, we formulated the two following sequence-based rules:

**Seq1** If two entities $E_1$ and $E_2$ are REVEALS/SEARCHES/TESTS-compatible and occur one after another with no other intervening entity, then a relation of this type should be inserted between $E_1$ and $E_2$. Figure 6 illustrates this rule.

**Seq2** If two entities $E_1$ and $E_2$ are MEANS-compatible and a third entity of type DOSE occurs between them, then a relation of type MEANS should be inserted between $E_1$ and $E_2$. This rule is illustrated by Figure 4.

***Nesting rules*** Recall (from Tab. 1) that the CAS corpus has a high rate of nested entities. In some cases, the precise types of the nesting and nested entity are a sufficient evidence of a relation existing between the two. For instance, in Figure 6 the entity [*échographie de l'appareil urinaire*]$^{\text{EXAMEN}}$ '[ultrasonogram of the urinary system]$^{\text{EXAM}}$' includes [*appareil urinaire*]$^{\text{ANATOMIE}}$ '[urinary system]$^{\text{ANATOMY}}$' and both are connected by the LOCATION relation. We experimentally found that this principle is quite accurate for LOCATION-compatible entities, which yields the following nesting-based rule:

**Nest1** If two entities $E_1$ and $E_2$ are LOCATION-compatible, $E_2$ is of type ANATOMY and is included in $E_1$, then a relation of this type should be inserted between $E_1$ and $E_2$. Figure 6 illustrates this rule.

The final relation extraction system consists in applying all the rules formalized above (Syn1, Syn2, Syn3, Lex1, Seq1, Seq2 and Nest1) and retaining all the relations inserted by them. The order of the rules does not matter for the final outcome. If the same relation was inserted by more than one rule, only one of its occurrences is retained. The following section describes the evaluation of this system on the test corpus.

## 6 Experimental results

As discussed in Sec. 3, we evaluate our approach on a test corpus composed of 5 files with 548 entities and 230 relations. The corpus contains not only sentences, but also text representation of tables. The latter was removed from the corpus beforehand, because our rules are not designed for it and, thus, false negatives can easily be generated.

The system was evaluated using a confusion matrix of the *true positives*, TP (i.e. relations correctly extracted), the *false positives* FP (i.e. relations extracted in wrong places) and the *false negatives*, FN (i.e. missed relations). The results are given in Table 4 along with precision, recall and F1-score. We notice an overall relatively high F1 score of 0.89. The number of false positives is low compared to the false negatives, indicating that our rules are specific and fail to cover many potential relations. This is mainly due to the selection of unambiguous relations. For example, the precision of 1 for LOCATION is due to nested entities in which this relation is sure to occur.

| Relation | TP | FN | FP | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| MEASURE_OF | 38 | 8 | 4 | 0,90 | 0,83 | 0.86 |
| MEANS | 23 | 7 | 2 | 0.92 | 0.76 | 0.84 |
| ACCOMPANIES | 5 | 2 | 1 | 0.83 | 0.71 | 0.77 |
| REVEALS | 52 | 3 | 3 | 0.95 | 0.95 | 0.95 |
| LOCATION | 65 | 12 | 0 | 1.00 | 0.84 | 0.92 |
| Total | 183 | 32 | 10 | 0.95 | 0.84 | 0.89 |

Table 4: Relation extraction results

An analysis of the texts annotated by the system, reveals that some errors came from wrong segmentation of text into sentences and words. However, most errors are due to incorrect dependency trees. As the majority of the relations are extracted using syntactic rules, the system is sensible to the syntactic variability. For example, in *La dose totale reçue lors de cette [perfusion]*$^{\text{MODE}}$ *a été égale à [30 mg]*$^{\text{DOSE}}$*"* 'The total dose received during this [perfusion]$^{\text{MODE}}$ was equal to [30 mg]$^{\text{DOSE}}$' there is no direct dependency between [*perfusion*]$^{\text{MODE}}$ and [*30 mg*]$^{\text{DOSE}}$, while there is one in the synonymous phrase *[perfusion]*$^{\text{MODE}}$ *de [30 mg]*$^{\text{DOSE}}$*" '[perfusion]*$^{\text{MODE}}$ *of [30 mg]*$^{\text{DOSE}}$*'*.

As we are exploiting syntactic rules, we cannot cover all patterns and may also need a larger corpus to identify other instances of the relations. After analysing the test corpus, we discover new relation patterns, only present in the test corpus. Nevertheless, the medical field remaining very specific, it limits this variability and makes it possible to obtain good results with few rules.

## 7 Conclusions

We have shown initial experiments and results in extracting semantic relations between entities in a medical corpus. A generic syntactic parser applied to a specialized text proved accurate enough to approximate semantic relations via syntactic dependencies. We also strongly exploited the fact that the relation between two entities is fully determined by the types of these entities. This allowed the rules to cover many relations at the same time. Another opportunity is the specificity of the text genre under study (clinical cases), in which dedicated lexico-syntactic constructions and entity sequences often repeat and can yield targeted rules.

Future work includes extending the sets of rules and experiments to larger corpora (e.g. by annotating relations in the whole CAS corpus). The resulting annotations might

then be used to train a model in a supervised setting. We will also further investigate the interface between information extraction from text on the one hand, and designing the schema of a knowledge graph and populating it from text on the other hand.

*Acknowledgements.* Work partly supported by the ICVL federation and RTR-DIAMS. It is done in the context of DOING action of the GDR-MADICS.

## References

1. Abacha, A.B., Zweigenbaum, P.: Automatic extraction of semantic relations between medical entities: a rule based approach. Journal of Biomedical Semantics **2**(Suppl 5), S4+ (2011)
2. Amavi, J., Halfeld Ferrari, M., Hiot, N.: Natural language querying system through entity enrichment. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium - International Workshop DOING, Lyon, France, August 25-27, 2020, Proceedings. Communications in Computer and Information Science, vol. 1260, pp. 36–48. Springer (2020)
3. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Res. **32**(Database-Issue), 267–270 (2004)
4. Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.L., Névéol, A.: A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Language Resources and Evaluation **52**(2), 571–601 (2017). https://doi.org/10.1007/s10579-017-9382-y, https://hal.archives-ouvertes.fr/hal-01631743
5. Cardon, R., Grabar, N., Grouin, C., Hamon, T.: Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In: Cardon, R., Grabar, N., Grouin, C., Hamon, T. (eds.) 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes. pp. 1–13. ATALA, Nancy, France (2020), https://hal.archives-ouvertes.fr/hal-02784737
6. Culotta, A., Sorensen, J.: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 423–429 (2004)
7. Embarek, M., Ferret, O.: Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco (2008)
8. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. Association for Computational Linguistics, Edinburgh, Scotland, UK. (Jul 2011), https://aclanthology.org/D11-1142
9. Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., Taylor, A.: Cypher: An evolving query language for property graphs. In: Das, G., Jermaine, C.M., Bernstein, P.A. (eds.) Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018. pp. 1433–1445. ACM (2018)
10. Franciscus, N., Ren, X., Stantic, B.: Dependency graph for short text extraction and summarization. J. Inf. Telecommun. **3**(4), 413–429 (2019)
11. Fundel, K., Küffner, R., Zimmer, R.: RelEx–Relation extraction using dependency parse trees. Bioinformatics **23**, 365–371 (2007)

12. Grabar, N., Grouin, C., Hamon, T., Claveau, V.: Corpus annoté de cas cliniques en français. In: TALN 2019 - 26e Conference on Traitement Automatique des Langues Naturelles. pp. 1–14. Toulouse, France (Jul 2019), https://hal.archives-ouvertes.fr/hal-02391878
13. Grouin, C., Grabar, N., Illouz, G.: Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. In: Denis, P., Grabar, N., Fraisse, A., Cardon, R., Jacquemin, B., Kergosien, E., Balvet, A. (eds.) Traitement Automatique des Langues Naturelles. pp. 1–13. ATALA, Lille, France (2021), https://hal.archives-ouvertes.fr/hal-03265926
14. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics (1992), https://aclanthology.org/C92-2082
15. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
16. Li, Z., Yang, Z., Shen, C., Xu, J., Zhang, Y., Xu, H.: Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. BMC Med Inform Decis Mak 19 **22** (2019)
17. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. pp. 404–411. ACL (2004)
18. Minard, A.L., Ligozat, A.L., Grau, B.: Multi-class SVM for Relation Extraction from Clinical Reports. In: Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria. pp. 604–609 (2011)
19. Minard, A.L., Roques, A., Hiot, N., Halfeld Ferrari Alves, M., Savary, A.: DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées. In: Cardon, R., Grabar, N., Grouin, C., Hamon, T. (eds.) 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes. pp. 66–78. ATALA, Nancy, France (2020), https://hal.archives-ouvertes.fr/hal-02784743
20. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore (Aug 2009), https://aclanthology.org/P09-1113
21. Ramadier, L., Lafourcade, M.: Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques. In: TALN: Traitement Automatique des Langues Naturelles. jep-taln2016, Paris, France (Jul 2016), https://hal.archives-ouvertes.fr/hal-01382323
22. Rindflesch, T.C., Bean, C.A., Sneiderman, C.A.: Argument Identification for Arterial Branching Predications Asserted in Cardiac Catheterization Reports. In: AMIA Annu Symp Proc. pp. 704–708 (2000)
23. Uzuner, O., Mailoa, J., Ryan, R., Sibanda, T.: Semantic relations for problem-oriented medical records. Artificial Intelligence in Medicine **50**, 63–73 (2010)