SIBM at CLEF e-Health Evaluation Lab 2015

Lina F. Soualmia[1,2], Chloé Cabot[1], Badisse Dahamna[1], Stéfan J. Darmoni[1,2]
[1] Normandie Univ., SIBM – TIBS – LITIS EA 4108, Rouen University and Hospital, France
[2] French National Institute for Health, INSERM, LIMICS UMR–1142, France
{surname.name}@chu-rouen.fr

Abstract. In this paper, we report on our participation in the *clinical named entity recognition* task of the CLEF eHealth 2015 evaluation initiative *i.e.* to fully automatically identify clinically relevant entities in medical text in French. We address the task by using several biomedical knowledge organization systems (KOS) containing terms and their variations already in French or that we have partially translated in the context of existing projects. The extraction method is available online in the form a web-based service that requests the KOS to extract clinical concepts from Electronic Health Records. It is also available via a user-friendly interface developed for clinicians. Our system has not obtained good results in inexact matching against the gold standard. However, this first participation allowed us to analyze our system and method and will allow us to improve it.

Keywords: Information extraction; Bagging; Lexical semantics; Natural Language Processing; Information storage and retrieval; Vocabulary controlled; Systematized Nomenclature of Medicine; Medical Subject Headings; International Classification of Diseases; Unified Medical Language System.

1 Introduction

With the increasing development of Electronic Health Records (EHRs) in hospitals and healthcare institutions [1], the amount of clinical documents, such as discharge summaries, in electronic format is also growing [2]. The retrieval of such documents is important in clinical and research tasks such as cohort studies or decision support in personalized medicine, a medicine tailored to each patient by considering genomic and clinical contexts of individuals. Indeed, these clinical documents are not only important to clinicians in daily use but also valuable to researchers and administrators. EHRs generate large amount of data that offer new opportunities to gain insight into clinical care. Particularly, EHR repositories enable to compose patient cohorts for the study of clinical hypotheses, hard to test experimentally, such as for example individual variability in drug responses. However, to compose those cohorts, efficient and user-friendly information retrieval systems are needed. To improve the performance of these systems, it is mandatory to develop an automatic indexing system that gives as output the representative index of an EHR. The latter should be represented by clinical related terms even if the discharge summaries are composed by free terms.

Since 1995, the department of BioMedical Informatics of the Rouen University Hospital (SIBM; URL: www.cismef.org) is working on developing tools to access health knowledge (information retrieval and automatic indexing) in French [3-8]. SIBM is a multidisciplinary team composed by physicians, medical informaticians, computer scientists, R&D engineers, librarians, postdoctoral and PhD students (n=21). SIBM is part of the Computer Science, Information Processing, and Systems Laboratory (LITIS-EA 4108), in Rouen, Normandy, France. Until recently, SIBM is working on the evaluation of health information systems and information retrieval and indexing in EHR [9-10]. In this context, a user-friendly tool and a web-based service ECMT (Extracting Concepts with Multiple Terminologies) is developed. It has been included in several projects subsidized by the French national research agency [11-12]. To evaluate the precision of ECMT, SIBM participated for the first time to the CLEF eHealth evaluation initiative [13]. The main motivation in participating is to improve the functionalities of the tool. The clinical named entity recognition task is retained [14]. It aims to fully automatically identify clinically relevant entities in medical texts in French. ECMT uses natural language processing (NLP), patterns and exploit several biomedical knowledge organization systems (KOS).

The rest of paper is organized as follows. In Section 2 we present related work, in Section 3 we introduce our extraction approach and tool and we describe our experimental setup. Section 4 reports on our results and on error analysis and reflections. Finally, Section 5 wraps up concluding remarks and outlines future work.

2 Related Work

Information extraction is the extraction of pre-defined types of information from text [15]. There are four primary methods available to implement an information extraction system, including Natural Language Processing (NLP), pattern matching, rules, and machine learning. The primary disadvantage of machine learning used for information extraction is that it requires a labeled dataset for training [16]. As most clinical data are stored in free text, the primary means of performing information extraction is natural language processing [17]. Several NLP systems have shown promising results in extracting information from medical narratives [18-21]. In [22], Turchin et al. used regular expressions (a meta-language which describes string search patterns), to extract numeric data form free-text. The use of rules and pattern-matching exploits basic patterns over a variety of structures, such as text strings, part-of-speech tags, semantic pairs, and dictionary entries [23]. Patterns are easily recognized by humans and can be expressed directly using special purpose representation languages such as regular expressions. Regular expressions are effective when the structure of the text and the tokens are consistent, but tend to be one-off methods tailored to the extraction task. Regular expressions have been used to extract blood pressure values from progress notes [22]. NLP has been useful for extracting medical information such as principal diagnosis [20] and medication use [24] from clinical narratives.

Using tools built over ontologies or controlled vocabularies such as the Systematized NOmenclature of MEDicine-Clinical Terms (SNOMED-CT) or the International Classification of Diseases-10 (ICD-10) have enabled researchers to automate the cap-

ture of information in clinical narratives [20]. Other tools have been developed. For example Aronson et al. [25] developed the Medical Text Indexer. It is based on matching document terms with UMLS terms [26] using MetaMap, comparing the phrases of the document with the phrases of the concepts using the trigram method and extracting MeSH terms from the k-nearest neighbors (kNN) of the document to be indexed. The indexing method of Névéol et al. [5] combines a linguistic method and kNN. The EAGL method [27] combines the vector space model (VSM) and a regular expression pattern matcher. BioAnnotator [28] uses a parser to identify noun phrases from a document and then matches them to UMLS concepts using a rule engine. AMTEx (automatic MeSH term extraction) [29] applies the C/NC value method, which allows extraction of composed terms from the text combining statistic and linguistic information and ranks the terms according to the value of C/NC. Only terms belonging to MeSH terms are kept. Jonquet et al. [30] applied the Mgrep tool for extracting concepts using 200 biomedical ontologies and computed a score for each generated annotation according to its origin (preferred term, non-preferred term, synonym term…). BioDI [31] reduces the limitation of partial matching through filtering MeSH concepts, which are extracted using VSM. MaxMatcher+ [32] exploits the BM25 weight for ranking the concepts extracted using MaxMatcher [33], which annotates documents with only the most significant words in the UMLS Metathesaurus.

All these methods are based on the use of one ore several biomedical KOS which link health concepts and gives their associated terms, as well as their definition and code. Such a system may take the form of a terminology, thesaurus, controlled vocabulary, nomenclature, classification, taxonomy, ontology …etc. Indeed, KOS facilitates the indexing, coding and annotation of different kinds of documents. In the health domain, a great number of bio-terminological resources have been developed for different purposes (the content and structure depending on the purpose to be served). This proliferation of resources has made finding the correct concept increasingly complex when using multiple terminologies simultaneously. For example, the ICD-10 was designed for coding medical reports, the MeSH Thesaurus, for document indexing, the ATC Classification, for coding drugs, the SNOMED-CT, for semantic interoperability among EHRs, and the MedDRA for adverse drug events. However, few of these resources are available for languages other than English [34]. SIBM developed and maintains a Health Terminologies and Ontologies Portal (HeTOP) [35] that contains 55 KOS in several languages. ECMT relies on the information system of HeTOP.

3 Material &Method

3.1 Extracting Concepts with Multiple terminologies : ECMT

ECMT is developed to extract as accurately as possible from texts as input, a list of candidate health concepts from the 55 KOS included in HeTOP. The extraction is performed at the phrase level of the text. A SOAP and REST Web services allow to provide a response in XML for each concept and contains: the offset of the first and the final word contained in the health concept, and which led to a medical concept in

the final list, the identifier and its semantic type if the health concept is included in the UMLS Metathesaurus, and the medical specialty of the concept. The latter are based on manual semantic links between general medical specialties (e.g. dermatology, oncology …etc.) and the KOS included in HeTOP. ECMT relies on bag-of-words and also pattern-matching designed for discharge summaries, procedure reports or laboratory results which contains symbolic data (presence or absence), numerical data and units of measurement. The method of bag-of-words was developed mainly for information retrieval and it has been adapted for indexing i.e. only the largest set of words that maps a concept label is extracted, even if is subsets map other concepts. The method is considered as being more precise and avoiding noise. The text in input is normalized and each phrase is processed separately to extract the concepts.

ECMT has also a user-friendly interface (Fig. 1) accessible after authentication (http://ecmt.chu-rouen.fr/). Several options are available to index the text:

• "**c**" : categorizing. If "c=true" the specialties of each extracted concept are given as output and their UMLS semantic type (default value: "true").

• "**r**" : refined. If "r=true" the search is stopped when a concept that matches a maximum of words is extracted (default value: "true"). For example, for "*cardiopathie hypertensive*", if "r=true" only the concept "*hypertension artérielle*" is returned; if "r=false", the method returns "*hypertension artérielle*" and "*maladie cardiaque*" (the latter is returned because "*cardiopathie*" is a synonym of the concept "*maladie cardiaque*").

• "**sn**": semantic network. If "sn=true" the concepts that are related directly (aligned [36]) to the concepts of the text are also returned by ECMT; (default value: "false").

• "**e**": exclusions. It is a string containing the identifiers of concepts to exclude (a specialty, a semantic type, a broader or a narrower term…etc.). For example, "e=CIS_MT_8,UML_ST_T060, MSH_D_C,T_DESC_PHARMA_RACINE" returns only concepts that are not "chirurgies" (CIS_MT_8) nor "procedures de diagnostic" (UML_ST_T060) nor MeSH "Maladies" (MSH_D_C) nor "racines de spécialités pharmaceutiques" (T_DESC_PHARMA_RACINE) (default value:"", all the categories are returned, the user can filter them after the extraction). In the case of the use of a father concept, all its descendants are excluded in the output.

• "**fi**" : filters. It is a string containing the identifiers of concepts to keep in the output (same as "e").

• "**a**" : ancestors. If "a=true" ECMT returns also the ancestors of each concept (default value: "false").

• "**d**" : descendants. If "d=true" ECMT returns the descendants of each concept (default value: "false").

• "**at**" : alternative terms. If "at=true" the synonyms of the concepts are also returned in the output (default value: "true").

The answer of the web-based service is an XML file which serializes the output of the annotation of the text. The following tags compose it:

• <cis-sentences> : the set of phrases that correspond to the input.

• <timemillis> : processing time in ms.

• <cis-sentence> : a phrase.

• <idsentence> : identifier of the phrase.

- <position> : beginning position of the phrase in the text.
- <start> : beginning position of indexing.
- <end> : end position of indexing.
- <idterm> : concept identifier in the original KOS.
- <offset> : set of the terms positions composing the concept.
- <ter> : acronym of the concept KOS.
- <umlscui> : UMLS concept identifier.
- <matchterms> : set of labels that allowed to retrieve the concept.
- <cis:term> : preferred label of the concept.
- <cis:label> : label.
- <lang> : label language.
- <cis:altterms> : list of alternative labels of the concept.
- <cis:altterm> : alternative label of the concept (synonym).
- <cis:categorization> : list of specialties or semantic types.
- <cis:category> : a specialty or a semantic type.
- <cis:descendants> : list of all descendants of the concept.
- <cis:descendant> : a descendant of the concept.
- <cis:ancestors> : list of all ancestors of the concept.
- <cis:ancestor> : an ancestor of the concept.
- <cis:relateds> : list of all concepts related semantically with the concept.
- <cis:related> : a concept related semantically with the concept.
- <relationLabel> : label of the relation.

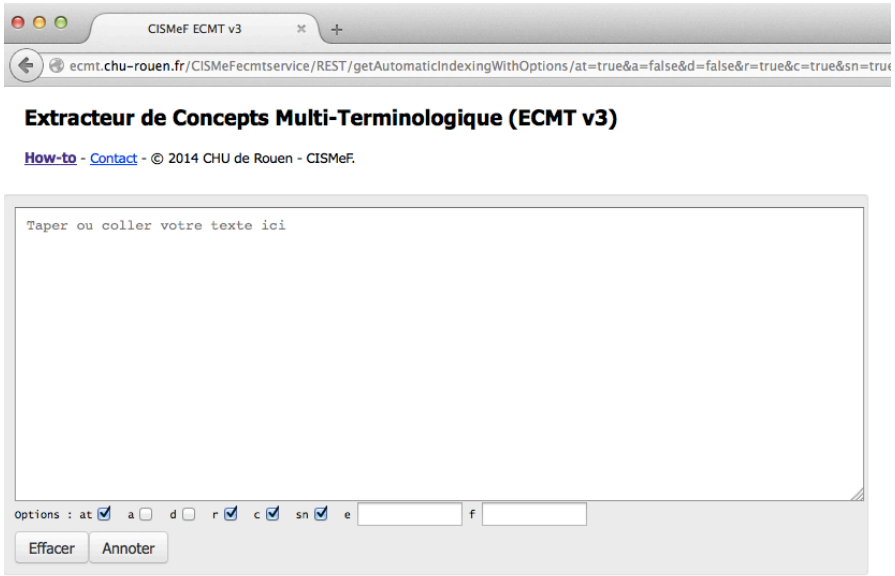

Fig.1. User interface and url of ECMT and its options. The default values are selected.

Fig.2. gives an an example of processing the phrase "*La contraception par les dispositifs intra utérins*". ECMT extracts the MeSH terms "*dispositifs contracptifs*" (CUI

C0009886), "*dispositifs intra-utérins*" (CUI C0021900) and the ATC term "contraceptifs intra-utérins" (CUI C3653534). The user can also visualize the alternative terms and categories (Fig.3).

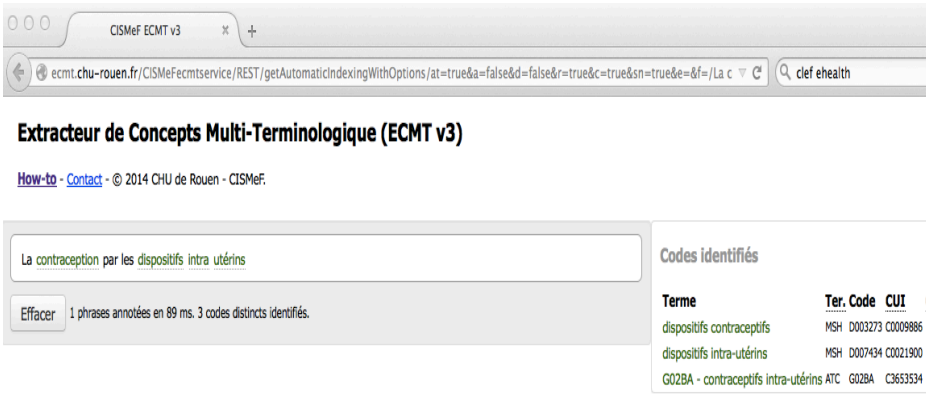

Fig.2. Example of processing the phrase *La contraception par les dispositifs intra utérins.*

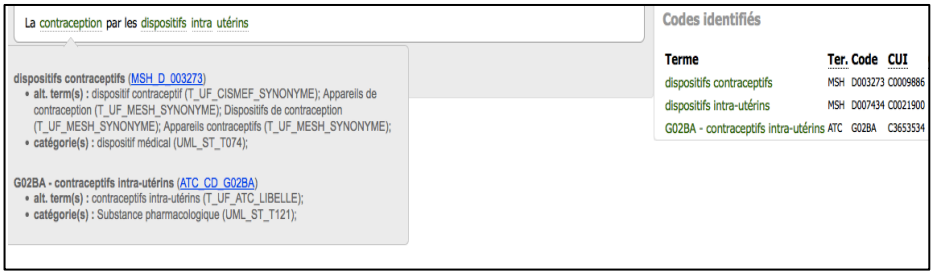

Fig.3. Visualization of alternative terms (synonyms) and categories (specialties or UMLS semantic types).

3.2 Biomedical Knowledge Organisation Systems

The information retrieval system of HeTOP, and thus of ECMT, operates on more than 55 terminologies in both French and English partially or totally translated into French, aligned with semantic relations. However, for the latest version of ECMT (v3), the relational database management system is replaced by the distributed cache Infinispan to allow fast processing of the inputs (the example of Fig.2 is processed in 89 ms). The main objectives are the optimization of the response times and the dissociation of the search engine from a proprietary RDBMS. The NoSQL solution Infinispan allows data distribution and calling from several web-based servers. The version with Hibernate search combined with Apache Lucene for full text indexing is retained. This configuration allows ECMT the processing of 70,000 electronic health records per day, using the 55 KOS.

At the date of the challenge of the CLEF-eHealth task 1b [14], seven KOS were migrated to Infinispan and were available for ECMT: the Medical Subject Headings, the Anatomical Therapeutic Chemical classification, the Classification Commune des

Actes Médicaux, the Classification Internationale des Maladies - 10ème révision, MedlinePlus, the Systematized Nomenclature of MEDicine International, and Pharmacology. Table 1 contains their metrics. Each concept of these KOS, when it is available in the UMLS, has a Concept Unique Identifier. It is the case for example for the CIM-10 and not for the CCAM.

Tab.1. Total of terms in French (preferred, concept labels, synonyms ..etc) of the KOS used in the task.

ATC	12,162
CCAM	25,621
CIM-10	107,940
MelinePlus	879
MeSH	289,457
SNOMeD-Int.	151,683
Pharma	34,261

3.3 Dataset

The data set is the QUAERO French Medical Corpus, which has been developed as a resource for named entity recognition and normalization in 2013 [37]. The data set has been created by Névéol et al. in the wake of the 2013 CLEF-ER challenge, with the purpose of creating a gold standard set of normalized entities for French biomedical text. A selection of the MEDLINE titles and EMEA documents used in the 2013 CLEF-ER challenge were selected for human annotation and are used in this challenge. Annotations are provided in the BRAT[1] standoff format and the annotation process was guided by concepts in the UMLS. Ten types of clinical entities which are UMLS Semantic Groups were annotated: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. The annotations were made in a comprehensive fashion, so that nested entities were marked, and entities could be mapped to more than one UMLS concept.

In particular: (i) If a mention can refer to more than one Semantic Group, all the relevant Semantic Groups should be annotated. For instance, the mention "*récidive*" (recurrence) in the phrase "*prévention des récidives*" (recurrence prevention) should be annotated with the category "DISORDER" (CUI C2825055) and the category "PHENOMENON" (CUI C0034897); (ii) If a mention can refer to more than one UMLS concept within the same Semantic Group, all the relevant concepts should be annotated. For instance, the mention "*maniaques*" (obsessive) in the phrase "*patients maniaques*" (obsessive patients) should be annotated with CUIs C0564408 and C0338831 (category "DISORDER"); (iii) Entities which span overlaps with that of another entity should still be annotated. For instance, in the phrase "*infarctus du myocarde*" (myocardial infarction), the mention "*myocarde*" (myocardium) should be annotated with category "ANATOMY" (CUI C0027061) and the mention "*infarctus du myocarde*" should be annotated with category "DISORDER" (CUI C0027051).

[1] http://brat.nlplab.org/standoff.html

4 Results & Discussion

For each run (MEDLINE and ELMA) the web-based service of ECMT is used. Before submitting our runs, we have tested ECMT with the default options (described in the section 3.1) and with the 7 available KOS for extracting *entities* and *normalized entities*. For the concerns of the task and the evaluation, the ECMT output is converted into the BRAT format. Fig.4. is the annotation file obtained and related to the phrase of Fig.2. *La contraception par les dispositifs intra utérins*.

```
T1      DEVI 3 36      dispositifs contraceptifs
#1      AnnotatorNotes T1     C0009886
T2      DEVI 25 50     dispositifs intra-utérins
#2      AnnotatorNotes T2     C0021900
T3      CHEM 3 50      G02BA - contraceptifs intra-utérins
#3      AnnotatorNotes T3     C3653534
```

Fig.4. Annotation file in BRAT containing entities and normalized entities extracted via ECMT

The results (inexact match) on the test on 400 files to extract entities of the training set are in Table 2. These first encouraging results obtained few days before the run submission deadline led us to participate to the challenge.

Tab.2. Results (inexact match) on 400 files (training set) and 7 KOS.

	TP	FP	FN	Precision	Recall	F1
ANAT	142	149	54	0.4880	0.7245	0.5832
CHEM	153	38	108	0.8010	0.5862	0.6770
DEVI	13	12	6	0.5200	0.6842	0.5909
DISO	375	96	209	0.7962	0.6421	0.7109
GEOG	14	4	7	0.7778	0.6667	0.7179
LIVB	125	38	31	0.7669	0.8013	0.7837
OBJC	3	16	28	0.1579	0.0968	0.1200
PHEN	14	35	17	0.2857	0.4516	0.3500
PHYS	60	33	74	0.6452	0.4478	0.5286
PROC	195	105	109	0.6500	0.6414	0.6457
Overall	**1094**	**526**	**643**	**0.6753**	**0.6298**	**0.6518**

The results obtained for the challenge (exact match precision, recall and F-score) are presented in tables 3, 5, 7 and 9 below (MEDLINE and EMEA) and are reported in [13]. We also present inexact performance scores in tables 4, 6, 8 and 10.

Tab.3. MEDLINE titles exact match overall : entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	680	2297	4412	0.22840	0.13350	0.16850
Average scores				0.35493	0.49746	0.39588
Median scores				0.38785	0.93750	0.45375

Tab.4. MEDLINE titles inexact match overall : entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	680	866	1205	0.70910	0.63660	0.67090
Average scores				0.52325	0.72405	0.57550
Median scores				0.58720	0.78970	0.66555

Tab.5. EMEA exact match overall : entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	9	2251	4124	0.00400	0.00220	0.00280
Average scores				0.30912	0.32842	0.31087
Median scores				0.21170	0.18355	0.22425

Tab.6. EMEA inexact match overall : entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	982	1278	2307	0.43450	0.29860	0.35390
Average scores				0.48158	0.51984	0.48808
Median scores				0.57675	0.55005	0.55385

Tab.7. MEDLINE titles exact match overall : normalized entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	1020	2434	4461	0.29530	0.18610	0.22830
Average scores				0.32138	0.42388	0.33632
Median scores				0.29530	0.40330	0.22830

Tab.8. MEDLINE titles inexact match overall : normalized entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	1993	1991	3485	0.50030	0.36380	0.42130
Average scores				0.42804	0.50526	0.45238
Median scores				0.50030	0.57350	0.42130

Tab.9. EMEA exact match overall : normalized entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	10	2255	4128	0.00440	0.00240	0.00310
Average scores				0.28546	0.27384	0.27928
Median scores				0.00440	0.00710	0.00470

Tab.10. EMEA inexact match overall : normalized entities.

	TP	FP	FN	Precision	Recall	F1
ECMT	596	1990	3542	0.23050	0.14400	0.17730
Average scores				0.42454	0.43218	0.42306
Median scores				0.42340	0.58170	0.49010

The results obtained for the challenge are not satisfactory at all, specifically for the EMEA corpus. The bad results obtained for the MEDLINE corpus should be explained by the existing doubloons in the KOS (Tab.11) that decrease the precision, and by the concepts extracted even if the KOS is not included in the UMLS, and thus no CUI and no semantic group are available in the output, giving noise. Also, the bad exact match results, compared to inexact match results, could be explained by slight differences in terms used. The gold standard uses UMLS labels while ECMT outputs preferred labels in the original KOS. This leads to minor differences between CLEF and ECMT outputs, such as *douleur* in CLEF output vs. *douleur***s** in ECMT output. Finally, as no specific processing was done to extract overlapping entities as described for the task [14], several nested entities are missed. For example, in Fig. 4. only the concept "C0021900" is in common with the gold standard (Fig.5). Other entities are extracted with ECMT but are not in the gold standard. As they are more precise, these concepts should not be considered as noise.

Tab.11. Total of terms (distinct) in French (preferred, concept labels, synonyms …etc) of the KOS used in the task.

ATC	11,322
CCAM	25,609
CIM-10	107,790
MelinePlus	877
MeSH	288,016
Pharma	34,172
SNOMeD-Int.	151,407

```
T1      PROC 3 16       contraception
#1      AnnotatorNotes T1 C0700589
T2      DEVI 25 50      dispositifs intra utérins
#2      AnnotatorNotes T2 C0021900
T3      ANAT 43 50 utérins
#3      AnnotatorNotes T3 C0042149
```

Fig.5. Annotation file in BRAT format : gold standard.

The results obtained for the EMEA corpus are null (Tab.5, Tab.6, Tab.9, Tab.10). These should be explained by the presence of specific characters in the text. Fig. 6 and Fig. 7 give an example the processing of an EMEA document excerpt: "*Dans quel cas Tysabri est-il utilisé ? Tysabri est utilisé dans le traitement des adultes atteints de sclérose en plaques*", all the rest of the phrase after the character "*?*" is ignored. Also, some characters such as "*:*" "μ" or newlines cause offsets to be shifted, due to specific ECMT processes, leading to decreased exact match results, especially in EMEA documents which contain many of those characters.

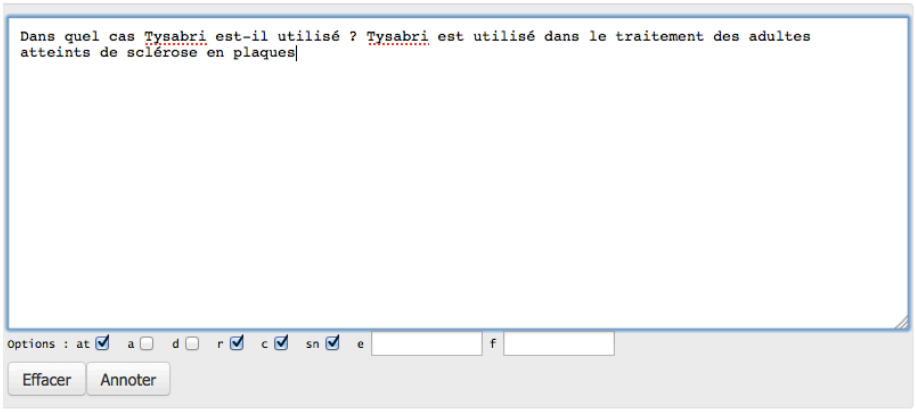

Fig.6. Testing ECMT with an excerpt of EMEA document**.**

Fig.7. Result of testing ECMT with an excerpt of EMEA document : no CUI is given in the output and the text after the character "?" is ignored**.**

After the submission of the runs of ECMT, the migrating process of the 55 KOS of HeTOP into Infinispan was achieved. A set of 32 KOS in French are available (Tab.12). We have tested all the training dataset (832 vs. 400 in the first test) by using the initial 7 KOS used in the challenge and also the 32 ones.

The obtained results are reported in tables 13, 14, 15 and 16 hereafter. The results are not null but neither satisfactory. Including several KOS increases the precision and decreases the recall in exact matching.

Tab.12. The 32 KOS in French included in the data grid of Infinispan. Number of distinct terms in French.

Adicap	8,721
ATC	11,322
BNCP	803
CCAM	25,609
CGP	220
CIF	1,503
ICD	105,790
CIO	1,603
CIP	1,240
CIS	2,169
Cladimed	4,672
FMA	26,906
GO	551
HPO	13,942
ICN	2,819
LNC	65,612
LPP	4,682
MedDRA	72,628
MED	877
MIM	361
MeSH	288,016
NAB	1,052
NCIT	51,414
Orphan	20,136
PAS	6,026
Pharma	34,172
Radlex	7,730
SNOMeD-CT	140,237
SMD	929
SNOMeD-Int	151,407
UMLS ST	147

Tab.13. Results (832 files) for Exact match ECMT using 7 KOS and default options.

	TP	FP	FN	Precision	Recall	F1
ANAT	80	411	235	0.1629	0.2524	0.1980
CHEM	86	258	736	0.2500	0.1046	0.1475
DEVI	7	32	28	0.1795	0.2000	0.1892
DISO	225	725	1143	0.2368	0.1645	0.1941
GEOG	21	12	17	0.6364	0.5526	0.5915
LIVB	89	205	182	0.3027	0.3284	0.3150
OBJC	2	25	44	0.0741	0.0435	0.0548
PHEN	9	51	50	0.1500	0.1525	0.1513
PHYS	42	118	232	0.2625	0.1533	0.1935
PROC	86	486	448	0.1503	0.1610	0.1555
Overall	**647**	**2323**	**3117**	**0.2178**	**0.1719**	**0.1922**

Tab.14. Results (832 files) for exact match using 32 KOS in French and default options.

	TP	FP	FN	Precision	Recall	F1
ANAT	90	401	498	0.1833	0.1531	0.1668
CHEM	98	246	1064	0.2849	0.0843	0.1301
DEVI	7	32	56	0.1795	0.1111	0.1373
DISO	276	674	1782	0.2905	0.1341	0.1835
GEOG	23	10	31	0.6970	0.4259	0.5287
LIVB	94	200	381	0.3197	0.1979	0.2445
OBJC	4	23	117	0.1481	0.0331	0.0541
PHEN	11	49	130	0.1833	0.0780	0.1095
PHYS	41	119	441	0.2563	0.0851	0.1277
PROC	146	426	860	0.2552	0.1451	0.1850
Overall	**790**	**2180**	**5360**	**0.2660**	**0.1285**	**0.1732**

Tab.15. Results (832 files) for inexact match using 7 KOS and default options.

	TP	FP	FN	Precision	Recall	F1
ANAT	250	241	86	0.5092	0.7440	0.6046
CHEM	254	90	169	0.7384	0.6005	0.6623
DEVI	21	18	13	0.5385	0.6176	0.5753
DISO	778	172	331	0.8189	0.7075	0.7557
GEOG	27	6	12	0.8182	0.6923	0.7500
LIVB	199	95	65	0.6769	0.7538	0.7133
OBJC	6	21	37	0.2222	0.1395	0.1714
PHEN	15	45	42	0.2500	0.2632	0.2564
PHYS	101	59	144	0.6313	0.4122	0.4988
PROC	390	182	134	0.6818	0.7443	0.7117
Overall	**2041**	**929**	**1033**	**0.6872**	**0.6640**	**0.6754**

Tab.16. Results (832 files) for inexact match using 32 KOS and default options.

	TP	FP	FN	Precision	Recall	F1
ANAT	225	266	155	0.4582	0.5921	0.5166
CHEM	250	94	211	0.7267	0.5423	0.6211
DEVI	22	17	29	0.5641	0.4314	0.4889
DISO	784	166	537	0.8253	0.5935	0.6904
GEOG	28	5	25	0.8485	0.5283	0.6512
LIVB	196	98	153	0.6667	0.5616	0.6096
OBJC	12	15	105	0.4444	0.1026	0.1667
PHEN	19	41	117	0.3167	0.1397	0.1939
PHYS	93	67	305	0.5813	0.2337	0.3333
PROC	405	167	307	0.7080	0.5688	0.6308
Overall	**2034**	**936**	**1944**	**0.6848**	**0.5113**	**0.5855**

5 Perspectives for future work

SIBM participated for the first time to an evaluation challenge. The *clinical named entity recognition* task of the CLEF eHealth 2015 evaluation initiative [13] allowed us

to evaluate ECMT in a very specific context (indexing MEDLINE titles and EMEA documents in French). ECMT is developed to index Electronic Health Records via a web-based service and also via a user-friendly interface. The actual version of ECMT (v3) is optimized to process around 70,000 EHR per day. ECMT was not trained with the training sets of the challenge and it used the default options and the 7 (vs. 55 today) KOS. For this kind of challenge, *clinical* named entity recognition, it would be more interesting, in our point of view, having a dataset *clinical* documents in French instead MEDLINE titles or EMEA documents with special characters.

The main conclusion of this work and the obtained results is that before running the datasets we should have studied the training sets and identified for example the specialized characters that are ignored by ECMT (mainly in the EMEA corpus). We should have also identified the set of KOS that gives the best results. We should have also tested the combinations of the options vs. the default values. For instance, for managing overlapping entities, the value of "`r`" should be `r=false` to avoid the recognition of only the concept that maps the largest bag-of-words. For normalized entities, the value of the parameter "`sn`" should be `sn=true` to exploit all the existing mappings until recognizing an UMLS concept that belongs to the semantic groups of the task. We expect doing this tuning parameter in the near future. We project to participate to other similar challenges but with a better training.

References

1. Jha AK, DesRoches CM, Kralovec PD, Joshi MS. A progress report on electronic health records in US hospital. *Health affairs* 2010, 29(10):1951-57.
2. Schuemie MJ, Sen E, Jong GW, Van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiology and drug safety* 2012, 21(6):651-8.
3. Darmoni SJ, Thirion B, Leroy JP, Douyère M, Lacoste B, Godard C, Rigolle I, Brisou M, Videau S, Goupy E, Piot J, Quéré M, Ouazir S, Abdulrab H. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the internet. *Medical Informatics and the Internet in Medicine* 2001, 26(3): 165-178.
4. Soualmia LF, Darmoni SJ. Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. International Journal of Medical Informatics 2005, 74(2-4):141-50.
5. Névéol A, Rogozan A, Darmoni SJ. Automatic indexing of online health resources for a French quality controlled gateway. *Information Processing & Management* 2006, 42(3) : 695-709.
6. Soualmia LF, Sakji S, Letord C, Rollin L, Massari P, Darmoni SJ. Improving information retrieval with multiple health terminologies in a quality-controlled gateway. *BMC Health Information Science and Systems* 2013, 1:8.
7. Griffon N, Schuers M, Soualmia LF, Grosjean J, Kerdelhué G, Kergoulay I, Dahama B, Darmoni SJ. A Search Engine to Access PubMed Monolingual Subsets: Proof of Concept - Evaluation in French. *Journal of Medical Internet Research* 2014, 16(12) : e271.
8. Chebil W, Soualmia LF, Omri MN, Darmoni, SJ. Indexing biomedical documents with a possibilistic network. *Journal of the Association for Information Science and Technology* 2015, in press.
9. Cabot C, Grosjean J, Lelong R, Lefebvre A, Lecroq T, Soualmia LF, Darmoni, SJ. Omic Data Modelling for Information Retrieval. *Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO,* 2014, pp. 415-424.

10. Lelong R, Merabti T, Grosjean J, et al. Moteur de recherche sémantique au sein du dossier du patient informatisé : langage de requêtes spécifique. *In proceeding of 15èmes Journées Francophones d'Informatique Médicale*, 2014, CEUR Workshop Proceedings Vol : 1323.
11. Dupuch M, Segond F, Bittar A, Dini L, Soualmia LF, Darmoni SJ, Gicquel Q, Metzger MH. Separate the grain from the chaff: make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records. *In proceedings of the 6th Language & Technology Conference*, 2013.
12. Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, Campillo B, Jouini W, Grosjean J, Massari P, Griffon N, Dupuch M, Tayalati F, Dugas E, Balvet A, Grabar N, Pereira S, Frandji B, Darmoni SJ, Cuggia M. RAVEL: Retrieval And Visualization in ELectronic health records. *In Studies in Health Technologies and Informatics*, 2012, 180:194-8.
13. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, Palotti J, Zuccon G. *Overview of the CLEF eHealth Evaluation Lab 2015*. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2015.
14. Névéol A, Grouin C, Tannier X, Hamon T, Kelly L, Goeuriot L, Zweigenbaum P. *CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition*. In CLEF 2015 Online Working Notes. CEUR-WS.
15. DeJong G. An overview of the FRUMP system. *Strategies for natural language processing*. 1982:149–176 (Chapter 5).
16. Zweigenbaum, P, Lavergne T, Grabar N, Hamon T, Rosset S, Grouin C. Combining an expert-based medical entity recognizer to a machine-learning system: methods and a case study. *Biomedical Informatics Insights*, 2013, 6(Suppl 1):51-62.
17. Hayes PJ, Carbonell J. Natural Language Understanding. *Encyclopedia of Artificial Intelligence* 1987:660–677.
18. Tange, H.J, de Hasman, PF, Schouten HC. Medical narratives in electronic medical records. *International Journal of Medical Informatics,* 1997, 46:7-29.
19. Taira, R. K., Soderland SG. A statistical natural language processor for medical reports. *Proceedings of the American Medical Informatics Association Symposium*, 1999: 970-4.
20. Zeng, Qing T, Goryachev S, Weiss S, Sordo M, Murphy SN, Ross L. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 2006 6:30.
21. Voorham J, Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners. *Journal of the American Medical Informatics Association*, 2007, 14(3):349-54.
22. Turchin A, Kolatkar NS, Grant RW, Makhni ML, Pendergrass EC, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association,* 2006, 13: 691-5.
23. Pakhomov S, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. *In proceedings of the Association for Computational Linguistics* 2005, 25–8.
24. Xu H, Stenner S, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association* 2010, 17:19–24.
25. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. *Medical Health Informatics*, 2004, 11(1): 268–272.
26. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 2004, 32(4): 267–270.
27. Ruch P. Automatic assignment of biomedical categories: Toward a generic approach. *Bioinformatics* 2006, 22(6): 658–664.

28. Mukherjea S, Subramaniam SV, Chanda G, Sankararaman S, Kothari R, Batra V, Bhardwaj D, Srivastava B. Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. *IBM Journal of Research and Development* 2004, 48(5–6): 693–701.
29. Hliaoutakis A, Zervanou K, Petrakis EGM. The AMTEx approach in the medical document indexing and retrieval application. *Data and Knowledge Engigneering* 2009, 68(3): 380–392.
30. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MF, Shah NH. NCBO resource index: Ontology-based search and mining of biomedical resources. *Journal of Web Semantics* 2011, 9(3): 316– 324.
31. Chebil W, Soualmia LF, Darmoni, SJ. BioDI: a new approach to improve biomedical documents indexing. *Proceedings of the 24th International Conference on Database and Expert Systems Applications* 2013: 78–87.
32. Dinh D, Tamine L. Towards a context sensitive approach to searching information based on domain specific knowledge sources. *Web Semantics: Science, Services and Agents on the World Wide Web* 2012, 12–13: 41–52.
33. Zhou X, Zhang X, Hu X. MaxMatcher: Biological concept extraction using approximate dictionary lookup. *In Pacific Rim International Conferences on Artificial Intelligence* 2006: 145–149.
34. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P. Language Resources for French in the Biomedical Domain. *Language and Resource Evaluation Conference*, 2014: 2146-2151.
35. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, Darmoni SJ. Health Multi-Terminology Portal: a semantics added-value for patient safety. *Studies in Health Technology and Informatics* 2011, Vol. 166: 129-138.
36. Merabti T, Soualmia LF, Grosjean J, Joubert M, Darmoni SJ. Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. Chapter in *Medical Informatics*, 2012 : 41-68. InTech Publishing.
37. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. *Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM* 2014:24-30.