SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND

Chloé Cabot[1], Lina F. Soualmia[1,2], Badisse Dahamna[1], and 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

Abstract. This paper presents SIBM's participation in the Multilingual Information Extraction task 2 of the CLEF eHealth 2016 evaluation initiative which focuses on named entity recognition in French written text. We report on the indexing of the provided QUAERO dataset with multiple knowledge organization systems (KOS) partially or totally translated in French. 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. We addressed the identification of relevant clinical entities within the International Classification of Diseases version 10 in the CépiDC dataset with a system based on natural language processing and approximate string matching methods. The results obtained this year were rather satisfactory and attested significant progress, particularly in exact match recognition, since our last year's participation.

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

Since the amount of digital medical documents has widely expanded in the last twenty years, the information retrieval from such heterogeneous documents has become a significant challenge to address a large variety of tasks in clinical and biomedical research as well as personalized medicine. Since 1995, the department of BioMedical Informatics of the Rouen University Hospital (SIBM, URL: www.cismef.org) has been working on developing tools to access health knowledge (information retrieval and automatic indexing) in French [1–6]. More recently, our team has worked on the evaluation of health information systems and information retrieval and indexing in Electronic Health Records (EHRs) [7,

8]. 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 [9, 10]. To evaluate the precision of ECMT, our team participated in 2015 for the first time to the CLEF eHealth evaluation initiative [11], precisely to the clinical named entity recognition task 1b [12, 13]. The results obtained during this previous edition were not satisfactory, partially due to our late-joining participation without training. This year, based on 2015 results, we participated in the multilingual information extraction task 2 (phases 1 et 2) [14, 15]. It aims to fully automatically identify clinically relevant entities in medical texts in French with several types of documents: abstracts titles, documents about marketed drugs and death certificates. The main motivation in participating is to improve the functionalities of the tool and to determine the progress achieved since our last year's participation and our ability to address the issues detected then. ECMT uses natural language processing (NLP), patterns and exploit several biomedical knowledge organization systems (KOS).

The rest of the paper is organized as follows. In Section 3 we introduce our extraction approach and tools used in QUAERO and CépiDC tasks 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 Material and methods

2.1 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 the Health Terminology / Ontology Portal (HeTOP) [16]. 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 is based on manual semantic links between general medical specialties (e.g. dermatology, oncology, etc.) [17] 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 contain symbolic data (presence or absence), numerical data and units of measurement. The method of bag-of-words [2] was developed initially 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 the input is normalized and each phrase is processed separately to extract the concepts. ECMT has also a user-friendly interface (Figure 1) accessible after authentication (http://ecmt.chu-rouen.fr/). Several options are available to index the text and described in [13].

A new option named `prioritization` was added since 2015. It addresses the specific issue related to the noise generated by multiple-terminology indexation. If this option is active, ECMT returns only the concept from the most reliable terminology, according to its semantic type (default value: `false`). When n identical terms from several terminologies are retrieved, semantic types related to these terms are computed and the most relevant is determined using set-theoretic operations. Then, the most pertinent term is retained based on a classification of the HeTOP resources devised manually for each semantic type available in UMLS. For example, indexing the term "asthme" (asthma) with ECMT results with 7 concepts retrieved within 7 different resources: SNOMED-int, NCIT, MeSH, Medline Plus, HPO, ICD-10 and ICDC. With the prioritization option activated, only one concept is retrieved according to the semantic type corresponding with "asthme" (T47-disease in this case) which is a MeSH concept. If no MeSH concept could be retrieved for a T47-disease concept, then an NCIT concept should be prioritized and retrieved, and so on. At this time, only 29 over 128 existing semantic types can be processed with this option.

Figure 2 gives an example of processing the phrase *Cholestases intrahépatiques fibrogènes familiales et anomalies héréditaires du métabolisme hépatocytaire des acides biliaires* with all ECMT default options but the activated prioritization option. ECMT extracts the MeSH terms *acides et sels biliaires* (CUI C0005391), *cholestase intrahépatique* (CUI C0008372), the ICD-10 term *E70-E90 anomalies du métabolisme* and the NCI term *héréditaire* (CUI C0439660). The user can also visualize the alternative terms and categories.

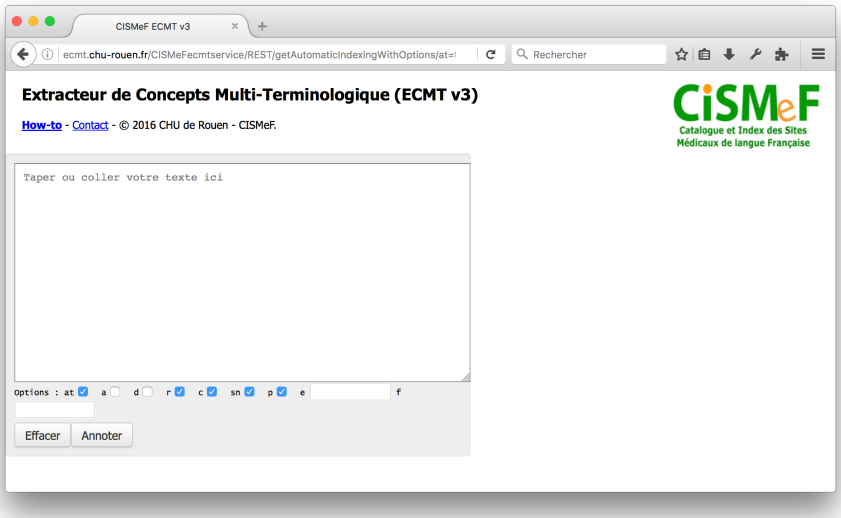

Fig. 1. User interface and URL of ECMT and its options. The default values are selected.

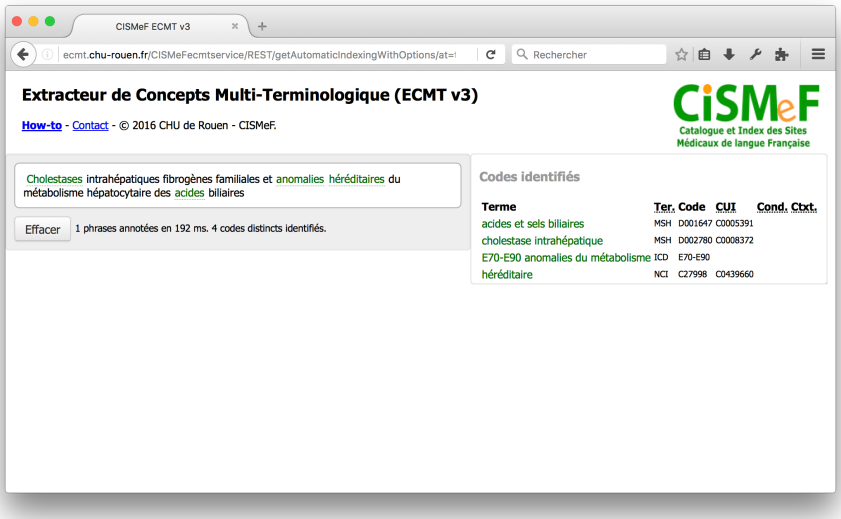

Fig. 2. Example of processing the sentence *Cholestases intrahépatiques fibrogènes familiales et anomalies héréditaires du métabolisme hépatocytaire des acides biliaires.*

2.2 Extracting Concepts from Death Certificates with ICD-10: CIMIND

The CépiDC track aims at identifying only ICD-10 terms with several versions of this resource manually curated by CépiDC (*see* section 2.4). This dataset made of death certificates revealed that many of the raw texts provided included spelling mistakes (french accents, inversions etc.). As ECMT is designed to perform only exact match using multiple terminologies, poor results have been obtained while analyzing the CépiDC corpus during the training phase. In this way, we choose to build CIMIND especially for the CépiDC track to focus on these particular issues.

CIMIND is designed to match ICD-10 concepts from the texts as input to ICD-10 terms in the relevant version of the ICD-10. The extraction is performed at the phrase level of the text using natural language processing techniques. The system is built using Python and Python/C extensions and provides a response in CSV format for each identified concept with: (i) the entry text, (ii) the offset of the first and the final word contained in the health concept, (iii) the ICD-10 identifier and (iv) the ICD-10 term. CIMIND performs three main steps to identify ICD-10 concepts:

Tokenization The input text is sliced into phrases, then words. Afterwards, stop words are filtered. Finally, spell checking is performed using the Enchant library.

The Enchant library is a generic spell checking library with a C API providing dictionaries and corrections for a misspelled word.

Candidate selection To select ICD-10 term candidates eventually matching the input phrase, a method based on the phonetic encoding algorithm Double Metaphone (DM) [18] is used to operate a first approximate term search. In this way, our system relies on a database storing pre-computed DM codes for each word available in each ICD-10 version dictionary. First, CIMIND computes DM codes for each word included in the analyzed phrase. Then, ICD-10 candidates with corresponding DM codes are retrieved from this database. This step provides quickly a list of relevant ICD-10 term candidates and allows us to perform time-consuming analyses on a reduced set of terms in the final step.

Candidate ranking Finally, a combination of the longest common substring and Levenshtein distance algorithms provides the candidate ranking. The most likely term having the highest score is retained as the matching ICD-10 term.

Figure 3 gives an example of processing the phrase *HEMATOME INTRACEREBRAL AVEC OEDEME ET ENGAGEMENT SOUSFALCIFORME* with CIMIND. CIMIND extracts the ICD-10 concepts *engagement sous-falciforme* (G935), *hématome intracérébral* (I619), and *oedème* (R609).

```
82944;2013;1;85;2;1;HEMATOME INTRACEREBRAL AVEC OEDEME ET ENGAGEMENT
SOUSFALCIFORME;NULL;NULL;;engagement sous-falciforme;G935
82944;2013;1;85;2;1;HEMATOME INTRACEREBRAL AVEC OEDEME ET ENGAGEMENT
SOUSFALCIFORME;NULL;NULL;;hématome intracérébral;I619
82944;2013;1;85;2;1;HEMATOME INTRACEREBRAL AVEC OEDEME ET ENGAGEMENT
SOUSFALCIFORME;NULL;NULL;;oedème;R609
```

Fig. 3. Annotation file in CSV containing ICD-10 concepts extracted with CIMIND.

Regarding execution time, CIMIND is able to process a death certificate as provided in the CépiDC corpus in about 80ms.

2.3 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. At the date of the challenge of the CLEF-eHealth 2016 task 2, thirteen KOS were migrated to Infinispan, a distributed in-memory key/value data store with optional schema, and were available for ECMT: the Medical Subject Headings (MeSH), the Anatomical Therapeutic Chemical classification (ATC), the Classification Commune des Actes Médicaux (CCAM), the International Classification of Diseases version

10 (ICD-10), MedlinePlus, the Systematized Nomenclature of MEDicine International (SNOMED-Int), Pharmacology, the International Classification of Primary Care (ICPC), the Foundational Model of Anatomy Ontology (FMA), the Human Phenotype Ontology (HPO), the NCI Thesaurus (NCIT), the Online Mendelian Inheritance in Man compendium (OMIM) and the Human Rare Diseases Ontology (HRDO). 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 ICD-10 and not for the CCAM.

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

KOS	**Total of terms**
MeSH	214,684
SNOMED-Int	151,479
NCIT	62,416
ICD-10	35,419
FMA	27,412
CCAM	23,154
HRDO	20,160
HPO	13,962
ATC	11,473
OMIM	7,218
Pharma	6,083
ICPC	2,120
Medline Plus	878
SIBM	254
UMLS Semantic Types	131
UMLS Semantic Groups	16

2.4 Datasets

The QUAERO dataset The QUAERO French Medical Corpus dataset has been developed as a resource for named entity recognition and normalization in 2013 [19]. The data set has been created by Neveol 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[3] 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).

The CépiDC dataset Since 1968, the CépiDC, a French National Institute for Health and Medical Research (Inserm) laboratory, is dedicated to elaborate annually the national medical causes of death statistics in association with the French National Institute for Statistics and Economic Studies (Insee), the dissemination of the data and the studies and researches on the medical causes of death. These statistics are built from information from the certificate of death. The CépiDC team handles a database containing more than 18,000,000 death records [20]. The CépiDC task consists of extracting ICD-10 codes from the raw lines of death certificate text. The task is an information extraction task that relies on the text supplied to extract ICD-10 codes from the certificates, line by line. The dataset includes 65,843 death certificates processed by CépiDC over the period 2006-2012. The corpus is supplied in CSV format and each row contains twelve information fields associated with a raw line of text from an original death certificate. The output comprises the 9 input fields plus two text fields used to report evidence text supporting the ICD-10 code supplied in the twelfth, final field. The tenth field should contain the excerpt of the original text that supports the ICD code prediction. The dataset also includes four versions of a manually curated ICD-10 dictionary developed at CépiDC.

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

3 Results and discussion

3.1 QUAERO track

For each track, the MEDLINE abstract titles and EMEA documents, the web-based service of ECMT is used. Before submitting our runs, we have tested ECMT with the following options actives: `refined`, `categorizing`, `semantic network`, `prioritization` and with the 7 (run2) or 13 (run1) available KOS for extracting entities and normalized entities. Run 1 uses the following resources: ATC, CCAM, ICDC, FMA, HPO, IDC-10, Medline Plus, MeSH, NCIT, OMIM, HPO, Pharma, SNOMED-Int. Run 2 uses the following resources: ATC, CCAM, ICD-10, Medline Plus, MeSH, Pharma, SNOMED-Int. For the concerns of the task and the evaluation, the ECMT output is converted into the BRAT format. Figure 4 is an example of the annotation file obtained with the following sentence: *L' hyperplasie médullosurrénalienne: une étiologie rare de l' hypertension artérielle – rapport d' un cas.*

```
T1 DISO 3 35 hyperplasie médullosurrénalienne
#1 AnnotatorNotes T1 C0020507
T2 DISO 63 86 hypertension artérielle
#2 AnnotatorNotes T2 C0020538
T3 ANAT 76 86 artérielle
#3 AnnotatorNotes T3 C0003842
```

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

The results obtained for the challenge are presented in tables 2, 3, 4, 5 (phase 1 entities and normalized entities) and tables 6 and 7 (phase 2 normalization).

To support our discussion, the results obtained in CLEF eHealth 2015 are presented in table 8 [13].

Phase 1: entities and normalized entities The results obtained for the phase 1 challenge are rather satisfactory, especially with the entity recognition with the following results: in exact match processing, we obtain a precision of 0.5381 and a recall of 0.3784 (run1) with the EMEA corpus and a precision of 0.6407 and a recall of 0.4375 (run2) with the MEDLINE corpus. In inexact match processing, we obtain a precision of 0.649 and a recall of 0.4869 (run1) with the EMEA corpus and a precision of 0.7668 and a recall of 0.5865 (run2) with the MEDLINE corpus.

For normalized entities, in exact match processing, we obtain a precision of 0.38 and a recall of 0.2687 (run1) with the EMEA corpus and a precision of 0.4776 and a recall of 0.3271 (run2) with the MEDLINE corpus. In inexact

match processing, we obtain a precision of 0.4005 and a recall of 0.2842 (run1) with the EMEA corpus and a precision of 0.4974 and a recall of 0.3412 (run2) with the MEDLINE corpus.

As of last year, our results have been improved, especially in exact match entity recognition. For the MEDLINE track, we improved the precision in exact match entity recognition by 280% and the recall is improved by more than three times. Since we corrected the processing of special characters in documents and the computed offsets, we have been able to actually process the EMEA documents in exact match and improve our results in inexact match as 2015 F1 is 0.35390 and 2016 F1s are 0.5564 (run1) and 0.5233 (run2).

Indexing with multiple terminologies leads to having duplicate terms in the results that decrease the precision. This fact explains the differences that can be observed between run1 (13 terminologies) and run2 (7 terminologies). Compared to last year, this issue has been considered and a new option has been added in ECMT. This option `prioritization` allows to retain only the most pertinent terms when several terminologies add up a same term in the output, and therefore reduce the noise. This ranking is operated according to the term semantic types. For each semantic type, a list of the most pertinent terminologies to be uppermost retained has been devised manually. However, as of today, only 29 semantic types over 128 are processed. The noise introduced by using multiple terminologies could then be even more reduced in the future.

Also, some errors in 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. “douleurs” in ECMT output. Finally, as no specific processing was done to extract overlapping entities as described in the task, several nested entities are missed. 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.

Phase 2: normalization The results obtained for the phase 2 challenge which we participated for the first time are also rather satisfactory. We obtain the following results: in exact match processing, we obtain a precision of 0.6044 and a recall of 0.4626 (run2) with the EMEA corpus and a precision of 0.5936 and a recall of 0.515 (run1) with the MEDLINE corpus. In inexact match processing, we obtain a precision of 0.605 and a recall of 0.463 (run2) with the EMEA corpus and a precision of 0.5938 and a recall of 0.5153 (run2) with the MEDLINE corpus.

In this phase as in the normalized entities track in phase 1, most errors in CUIs retrieved are due to differences between our data and the gold standard’s. As we used up to 13 terminologies from various sources, and HeTOP does not track versions of these resources yet, most of these errors are related to the data sources and can also be related to alignments between these sources (and their different versions) and the UMLS.

Table 2. QUAERO Phase 1 (EMEA) - Entities.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	834	716	1370	0,5381	0,3784	0,4443	1006	544	1060	0,649	0,4869	0,5564
SIBM-run2	724	483	1480	0,5998	0,3285	0,4245	866	341	1237	0,7175	0,4118	0,5233
Average				0,525	0,4114	0,435				0,6377	0,5141	0,5423
Median				0,5998	0,3784	0,4443				0,7175	0,4808	0,5564

Table 3. QUAERO Phase 1 (EMEA) - Normalized entities.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	592	966	1611	0,38	0,2687	0,3148	626	937	1577	0,4005	0,2842	0,3324
SIBM-run2	467	735	1736	0,3885	0,212	0,2743	500	710	1703	0,4132	0,227	0,293
Average				0,4762	0,3215	0,3761				0,4968	0,4341	0,4405
Median				0,4466	0,2687	0,3148				0,4666	0,2842	0,3324

Table 4. QUAERO Phase 1 (MEDLINE) - Entities.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	1476	1258	1626	0,5399	0,4758	0,5058	1799	935	972	0,658	0,6492	0,6536
SIBM-run2	1357	761	1745	0,6407	0,4375	0,5199	1624	494	1145	0,7668	0,5865	0,6646
Average				0,503	0,4264	0,4455				0,6387	0,5707	0,5859
Median				0,6166	0,4375	0,4981				0,7394	0,5682	0,6422

Table 5. QUAERO Phase 1 (MEDLINE) - Normalized entities.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	1103	1638	1994	0,4024	0,3562	0,3779	1152	1589	1946	0,4203	0,3719	0,3946
SIBM-run2	1013	1108	2084	0,4776	0,3271	0,3883	1057	1068	2041	0,4974	0,3412	0,4047
Average				0,5006	0,376	0,4287				0,5181	0,4757	0,4917
Median				0,4927	0,3826	0,4308				0,506	0,3917	0,4416

Table 6. QUAERO Phase 2 (EMEA) - Normalization.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	1047	800	1156	0,5669	0,4753	0,517	1048	799	1155	0,5674	0,4757	0,5175
SIBM-run2	1019	667	1184	0,6044	0,4626	0,524	1020	666	1183	0,605	0,463	0,5246
Average				0,5507	0,4729	0,5073				0,5511	0,4732	0,5077
Median				0,5669	0,4753	0,517				0,5674	0,4757	0,5175

Table 7. QUAERO Phase 2 (MEDLINE) - Normalization.

	exact match, overall						inexact match, overall					
	TP	FP	FN	Pre.	Rec.	F1	TP	FP	FN	Pre.	Rec.	F1
SIBM-run1	1598	1094	1505	0,5936	0,515	0,5515	1599	1094	1504	0,5938	0,5153	0,5518
SIBM-run2	1450	978	1651	0,5972	0,4676	0,5245	1452	978	1649	0,5975	0,4682	0,525
Average				0,5551	0,4854	0,5167				0,5553	0,4857	0,517
Median				0,5936	0,4736	0,5245				0,5938	0,4736	0,525

Table 8. Summary of SIBM CLEF eHealth 2015 task 1b results.

Corpus	Track	Precision	Recall	F1
MEDLINE	entities, exact match	0.22840	0.13350	0.16850
	entities, inexact match	0.70910	0.63660	0.67090
	normalized entities, exact match	0.29530	0.18610	0.22830
	normalized entities, inexact match	0.50030	0.36380	0.42130
EMEA	entities, exact match	0.00400	0.00220	0.00280
	entities, inexact match	0.43450	0.29860	0.35390
	normalized entities, exact match	0.00440	0.00240	0.00310
	normalized entities, inexact match	0.23050	0.14400	0.17730

3.2 CépiDC track

CIMIND is used to analyze the CépiDC dataset and outputs the results in CSV format. The results obtained from this CépiDC track is presented in table 9. In this track, we obtained a precision of 0.6964 and a recall of 0.6634. The number of terms retrieved are rather decent, but comparing to results of other teams participating in this track, our error rate is not satisfactory. As the CIMIND system has been built expressly for the CLEF eHealth 2016 challenge, we lacked time to improve the final step performed by our system by testing more edit distances and combinations of these methods and then upgrade performances. In this way, it would be quite interesting to participate again in such a task in the future.

Table 9. CépiDC results.

	exact match, overall					
	TP	FP	FN	Precision	Recall	F1
SIBM-run1	72192	31480	36626	0,6964	0,6634	0,6795
Average				0,7878	0,6636	0,7185
Median				0,811	0,6554	0,6997

4 Conclusion and perspectives

For the second year, the multilingual information extraction task 2 of the CLEF eHealth 2016 evaluation initiative allowed us to evaluate ECMT in a very specific context (indexing MEDLINE titles and EMEA documents in French). ECMT is developed to index EHRs 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. Then, ECMT is not quite designed for the kind of datasets, abstract titles and EMEA documents, proposed in this challenge. Nevertheless, since our first participation in 2015, we have been able to improve ECMT performances thanks to the first evaluation which then revealed several issues related mainly to special characters and offsets computed by ECMT.

The main conclusion of this work and the obtained results is that improvements are still to be performed to reduce the noise related to multiple terminology-indexing as our different runs have revealed. Also, the recognition itself could still be enhanced. Moreover, this year's edition has revealed that version tracking of the resources available in HeTOP could be a major improvement for ECMT in the future. Regarding the CepiDC track, progress could have been achieved with more time and prior knowledge of the documents provided in the challenge. We plan on deepen these two approaches and to participate to other challenges in the future to keep track of our developments.

References

1. Darmoni, S., Thirion, B., Leroy, J., 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 **26**(3) (2001) 165–178
2. Soualmia, L.F., Darmoni, S.J.: Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. International Journal of Medical Informatics **74**(2) (2005) 141–150
3. Névéol, A., Rogozan, A., Darmoni, S.: Automatic indexing of online health resources for a french quality controlled gateway. Information processing & management **42**(3) (2006) 695–709
4. Soualmia, L.F., Sakji, S., Letord, C., Rollin, L., Massari, P., Darmoni, S.J.: Improving information retrieval with multiple health terminologies in a quality-controlled gateway. BMC Health Information Science and Systems **1**(1) (2013) 1–8
5. Griffon, N., Schuers, M., Soualmia, L.F., Grosjean, J., Kerdelhué, G., Kergourlay, I., Dahamna, B., Darmoni, S.J.: A search engine to access pubmed monolingual subsets: Proof of concept and evaluation in french. Journal of medical Internet research **16**(12) (2014)
6. Chebil, W., Soualmia, L.F., Omri, M.N., Darmoni, S.J.: Indexing biomedical documents with a possibilistic network. Journal of the Association for Information Science and Technology **67**(4) (2016) 928–941
7. Cabot, C., Grosjean, J., Lelong, R., Lefebvre, A., Lecroq, T., Soualmia, L.F., Darmoni, S.J.: Omic data modelling for information retrieval. In: IWBBIO, Citeseer (2014) 415–424
8. Lelong, R., Merabti, T., Grosjean, J., Joulakian, M., Griffon, N., Dahamna, B., Cuggia, M., Pereira, S., Grabar, N., Thiessard, F., et al.: Moteur de recherche sémantique au sein du dossier du patient informatisé: langage de requêtes spécifique. 15es Journées francophones d'informatique médicale (2014) 139–151
9. Dupuch, M., Segond, F., Bittar, A., Dini, L., Soualmia, L., Darmoni, S., Gicquel, Q., Metzger, M.: 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 and Technology Conference. (2013)
10. Thiessard, F., Mougin, F., Diallo, G., Jouhet, V., Cossin, S., Garcelon, N., Campillo-Gimenez, B., Jouini, W., Grosjean, J., Massari, P., et al.: Ravel: retrieval and visualization in electronic health records. In: MIE. (2012) 194–198
11. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer (2015) 429–443
12. 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: Proceedings of CLEF. (2015)
13. Soualmia, L.F., Cabot, C., Dahamna, B., Darmoni, S.J.: SIBM at CLEF eHealth evaluation lab 2015, CLEF (2015) Working Notes
14. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: CLEF 2016 - 7th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer, September, 2016.

15. Névéol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth Evaluation lab 2016. In: CLEF 2016 Evaluation Labs and Workshop Online Working Notes, CEUR-WS, September, 2016.
16. Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion, B., Soualmia, L.F., Darmoni, S.J.: Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform **166**(66) (2011) 129–138
17. Darmoni, S.J., Névéol, A., Renard, J.M., Gehanno, J.F., Soualmia, L.F., Dahamna, B., Thirion, B.: A medline categorization algorithm. BMC medical informatics and decision making **6**(1) (2006) 7
18. Philips, L.: The double metaphone search algorithm. C/C++ users journal **18**(6) (2000) 38–43
19. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The Quaero french medical corpus: A resource for medical entity recognition and normalization. In: Proc BioTextM, Reykjavik, Citeseer (2014)
20. Pavillon, G., Laurent, F.: Certification et codification des causes médicales de décès. Bulletin épidémiologique hebdomadaire **30**(31) (2003) 134–138