Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016

Aurélie Névéol[1], K. Bretonnel Cohen[1,2], Cyril Grouin[1], Thierry Hamon[1,3]
Thomas Lavergne[1,4], Liadh Kelly[5], Lorraine Goeuriot[6]
Grégoire Rey[7], Aude Robert[7], Xavier Tannier[1,4], and Pierre Zweigenbaum[1]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
firstname.lastname@limsi.fr
[2] University of Colorado, USA
[3] Université Paris Nord, Villetaneuse, France
[4] Univ. Paris-Sud, Orsay, France
[5] ADAPT Centre, Trinity College, Dublin, Ireland
liadh.kelly@tcd.ie
[6] Université Grenoble Alpes, Grenoble, France
lorraine.goeuriot@imag.fr
[7] INSERM-CépiDC, Paris, France
firstname.lastname@inserm.fr

Abstract. This paper reports on Task 2 of the 2016 CLEF eHealth evaluation lab which extended the previous information extraction tasks of ShARe/CLEF eHealth evaluation labs. The task continued with named entity recognition and normalization in French narratives, as offered in CLEF eHealth 2015. Named entity recognition involved ten types of entities including *disorders* that were defined according to Semantic Groups in the Unified Medical Language System® (UMLS®), which was also used for normalizing the entities. In addition, we introduced a large-scale classification task in French death certificates, which consisted of extracting causes of death as coded in the International Classification of Diseases, tenth revision (ICD10). Participant systems were evaluated against a blind reference standard of 832 titles of scientific articles indexed in MEDLINE, 4 drug monographs published by the European Medicines Agency (EMEA) and 27,850 death certificates using Precision, Recall and F-measure. In total, seven teams participated, including five in the entity recognition and normalization task, and five in the death certificate coding task. Three teams submitted their systems to our newly offered reproducibility track. For entity recognition, the highest performance was achieved on the EMEA corpus, with an overall F-measure of 0.702 for plain entities recognition and 0.529 for normalized entity recognition. For entity normalization, the highest performance was achieved on the MEDLINE corpus, with an overall F-measure of 0.552. For death certificate coding, the highest performance was 0.848 F-measure.

Keywords: Natural Language Processing; Named Entity Recognition, Entity Linking, Text Classification, UMLS, French, Biomedical Text

1 Introduction

This paper describes an investigation of information extraction and normalization (also called "entity linking") from French-language health documents. The methodology applied is the shared task model. In shared tasks, multiple groups agree on a "shared" task definition, a shared data set, and a shared evaluation metric. The idea is to allow evaluation of multiple approaches to a problem while minimizing avoidable differences related to the task definition, the data used, and the figure of merit applied [1, 2].

Over the past three years, CLEF eHealth offered challenges addressing several aspects of clinical information extraction (IE) including named entity recognition, normalization [3, 4] and attribute extraction [5]. Initially, the focus was on a widely studied type of corpus, namely written English clinical text [3, 5]. Starting in 2015, the lab's IE challenge evolved to address lesser studied corpora, including biomedical texts in a language other than English i.e., French [4]. This year, we continue to offer a shared task based on a large set of gold standard annotated corpora in French. In addition to named entity extraction and entity normalization already offered in 2015 [6], we introduced a coding task that required normalized entity extraction at the sentence level.

The significance of this work comes from the observation that challenges and shared tasks have had a significant role in advancing Natural Language Processing (NLP) research in the clinical and biomedical domains [7, 8], especially for the extraction of named entities of clinical interest [9–12], and entity normalization [11, 13–16].

One of the goals for this shared task is to foster the development of NLP tools for French in spite of the known discrepancies in language resources available for French and other languages in the biomedical domain, compared to English [17]. Findings of last year's lab were that while there was a sustained interest in addressing French from teams all over the world, results were very heterogenous depending on methods and resources used, as well as technical issues encountered [6]. This year's lab suggests increased maturity of the task as major technical problems are now tackled, performance increases, and reproducibility is introduced as an additional goal.

2 Material and Methods

In the CLEF eHealth 2016 Evaluation Lab Task 2, two datasets were used. The QUAERO French Medical corpus was used for named entity extraction and normalization. The CépiDC corpus was used for coding. Further details on the datasets, tasks and evaluation metrics are given below.

2.1 Datasets

The QUAERO French Medical corpus The QUAERO French Medical Corpus [18] was used for named entity extraction and normalization in CLEF

eHealth 2015 (task 1b) and CLEF eHealth 2016 (task 2). The dataset will be shared freely with the community after the challenge results have been announced. For a detailed description of the QUAERO corpus, we refer interested readers to the corpus website `http://quaerofrenchmed.limsi.fr/` and to the 2016 task 1b report [6], which include a detailed description of the annotation guidelines and excerpts of the corpus. Table 1 presents statistics for the specific sets provided to participants in CLEF eHealth 2016. The training set released in the CLEF eHealth 2016 Task 2 challenge corresponds to the training set provided in the CLEF eHealth 2015 Task 1b challenge, the development set corresponds to the test set provided in the CLEF eHealth 2015 Task 1b challenge, and the test set was previously unreleased. EMEA documents were divided into several files for readability through the BRAT interface.

Table 1. Descriptive statistics of the QUAERO French Medical Corpus

	EMEA			MEDLINE		
	Training	Development	Test	Training	Development	Test
Documents	3	3	4	833	832	833
Tokens	14,944	13,271	12,042	10,552	10,503	10,871
Entities	2,695	2,260	2,204	2,994	2,977	3,103
Unique Entities	923	756	658	2,296	2,288	2,390
Unique CUIs	648	523	474	1,860	1,848	1,909

The CépiDC corpus The CépiDC Corpus was provided by the French institute for health and medical research (INSERM) for the task of ICD10 coding in CLEF eHealth 2016 (task2). It consists of free text death certificates collected from physicians and hospitals in France over the period of 2006–2013.

Table 2 presents statistics for the specific sets provided to participants. The training set covered the 2006–2012 period, and the test set covered the 2013 period. This time-oriented construction of the datasets reflects the practical use case of coding death certificates, where historical data is available to train systems that can then be applied to current data to assist with new document curation.

CépiDC Dataset excerpts Death certificates are standardized documents filled by physicians to report the death of a patient. The content of the medical information reported in a death certificate and subsequent coding for public health statistics follows complex rules described in a document that was supplied to participants [19]. Table 3 presents an excerpt of the CépiDC corpus that illustrates the heterogeneity of the data that participants had to deal with. While some of the text lines were short and contained a term that could be directly linked to a single ICD10 code (e.g., "Détresse respiratoire"), other lines could

Table 2. Descriptive statistics of the CépiDC French Death Certificates Corpus

	Training (2006–2012)	Test (2013)
Documents	65,844	27,850
Lines	195,204	80,899
Tokens	1,176,994	496,649
Total ICD codes	266,808	110,869
Unique ICD codes	3,233	2,363

be run-on (e.g., "Maladie de Parkinson ..."), contain non-diacritized text (e.g., "DENUTRITION" missing the diacritic on the "E"), a mix of cases and diacritized text ("DEMENCE MIXTE EVOLUEE (stade sévère)"), abbreviations (e.g., "membre sup" instead of "membre supérieur") and so on.

Table 3. Two sample documents from the CépiDC French Death Certificates Corpus. English translations for each text line are provided in footnotes.

line	text	ICD codes
1	Arrêt cardio respiratoire[1]	R092
2	Détresse respiratoire[2]	J960
3	Amyotrophie spinale de type I[3]	G120
1	DENUTRITION DESHYDRATATION[4]	E46 E86
2	DEMENCE MIXTE EVOLUEE (stade sévère)[5]	F03
5	Maladie de Parkinson idiopathique[6]	G200 R600
5	Angioedème membres sup récent non exploré par TDM	
5	(à priori pas de cause médicamenteuse)	

2.2 Tasks

Named entity recognition (QUAERO Corpus). The task of named entity recognition consisted of analyzing plain text documents in order to mark the ten types of entities of clinical interest defined in the lab (Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures). Participants could mark either plain entities (i.e.,

[1] *Cardio-respiratory arrest*
[2] *Acute respiratory failure*
[3] *Type 1 spinal muscular atrophy*
[4] *Malnutrition dehydration*
[5] *Advanced mixed dementia (late stage)*
[6] *Idiopathic Parkinson's disease Recent angioedema of upper extremities w/o CT exploration (no known drug cause)*

mark the text mentions referring to an entity of interest) or normalized entities (i.e., supply UMLS Concept Unique Identifiers corresponding to the entities in addition to marking mentions).

Entity normalization (QUAERO corpus). The task of entity normalization consisted of mapping entities of clinical interest marked in biomedical text to a relevant UMLS CUI.

ICD10 coding (CépiDC corpus). The task of coding consisted of mapping sentences in the death certificates to one or more relevant codes from the International Classification of Diseases, tenth revision (ICD10).

Replication. The replication task invited lab participants to submit a system used to generate one or more of their submitted runs, along with instructions to install and use the system. Then, two of the organizers independently worked with the submitted material to replicate the results submitted by the teams as their official runs.

2.3 Evaluation metrics

System performance was assessed by the usual metrics of information extraction: precision (Formula 1), recall (Formula 2) and F-measure (Formula 3; specifically, we used β=1.) for named entity recognition and entity normalization.

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

$$\text{F-measure} = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{3}$$

Performance measures were computed at the document level and micro-averaged over the entire corpus. We determined system performance by comparing participating system outputs against reference standard annotations on the test set. For the QUAERO corpus, results were computed using the brateval program initially developed by Verspoor et al. [20], which we extended to cover the evaluation of normalized entities. For the CépiDC corpus, results were computed using a perl program. The evaluation tools were supplied to task participants along with the training data.

For **plain entity recognition**, an exact match (true positive) was counted when the system's entity type and span matched the reference. A false positive was counted if the system's entity type and span did not exactly match the reference.

For **normalized entity recognition**, an exact match (true positive) was counted when the system's entity type, span and CUIs matched the reference. Partial credits were given when only a subset of the expected CUIs were supplied by the system for a given entity.

For **entity normalization**, matches (true positives) were counted for each CUI supplied with an entity. As a result, if either the system or the reference supplied a list of CUIs associated with an entity, partial credit was awarded if the reference and system lists were not identical but a subset of the lists matched. However, system CUIs absent from the reference lists were counted as false positives.

For **coding**, matches (true positives) were counted for each ICD10 full code supplied that matched the reference for the associated document line.

The evaluation of the submissions to the **replication** task was essentially qualitative: we used a scoring grid to record the ease of installing and running the systems, the time spent to obtain results with the systems (analysts were committed to spend at most one working day - or 8 hours - to work with each system), and whether we managed to obtain the exact same results submitted as official runs.

3 Results and Discussion

Participating teams included between two and eight team members and resided in France (teams ERIC-ECSTRA, LIMSI, LITL and SIBM), the Netherlands (team Erasmus), Switzerland (Team BITEM) and Spain (Team UPF). Teams often comprised members with a variety of backgrounds and drew from computer science, informatics, statistics, information and library science, clinical practice. It can be noted that one team (LITL) participated in the challenge as a master-level class project.

For the plain entity recognition task, five teams submitted a total of 9 runs for each of the corpora, EMEA and MEDLINE (18 runs in total). For the normalized entity recognition task, three teams submitted a total of 5 runs for each of the corpora (10 runs in total). For the normalization task, two teams submitted a total of 3 runs for each of the corpora (6 runs in total). For the coding task, five teams submitted a total of 7 runs.

Three systems were submitted, allowing us to attempt replicating a total of seven runs.

3.1 Methods implemented in the participants' systems

Participants used a variety of methods, many of which relied on lexical sources (medical terminologies and ontologies). Interestingly, some of these knowledge-based methods relied on the training data supplied in the challenge as an additional knowledge source. Some groups relied on statistical machine translation to address the limitation of French coverage in the lexical sources available to them. For each corpus, 3 teams out of 5 solely relied on knowledge-based sources,

and did not use machine learning for the specific task of entity recognition and normalization. The knowledge resources were used in combination with string matching or indexing methods that were sometimes guided by linguistic principles to identify entities and concepts in the challenge corpus.

Machine-learning methods were still used by 2 teams out of 5 for each corpus. They relied on Conditional Random Fields (CRFs), Latent Dirichlet Analysis (LDA), Support Vector Machines (SVMs), and statistical information retrieval models. They often used lexical resources as features.

Participants who worked with the QUAERO and the CépiDC corpus did not use the exact same systems to address both corpora.

BITEM The BITEM team participated in the entity recognition and coding tasks [21] using a different method for each task. Entity recognition in the QUAERO corpus relied on a categorizer using the French UMLS to suggests a ranked list of candidate entities potentially denoted by each text unit in the corpus. Then, a second module anchored these candidates in the text, and normalized the entities that could be anchored. For the coding task in the CépiDC corpus, an ad hoc solution was developped based on pattern matching. This method prioritizes exact matches that fit the whole text. Failing that, the longest match is then selected.

ERIC-ECSTRA The ERIC-ECSTRA team participated in the coding subtask [22]. Their first run is based on the probabilistic topic model approach. It relies on a supervised extension of the LDA model, called Labeled-LDA, that builds on the latent topical structures to predict a category. The idea is that knowledge of document topics can help predict the associated outputs (here the ICD10 codes). Their second run is based on an SVM classifier with a bag-of-word data representation. Their results suggest that Labeled-LDA and SVM both achieve competitive results. It is interesting to note that one advantage of the LabeledLDA method is that the classifier results are easier to understand for humans.

Erasmus MC The Erasmus MC team participated in the entity recognition and the ICD-10 coding tasks [23]. For both tasks a dictionary-based approach was followed. For entity recognition and normalization, the system that had been developed for the same task in the CLEF eHealth 2015 challenge [24], was tuned on the 2016 training data. Briefly, a locally developed tagger, Peregrine, used a dictionary consisting of French terminologies from the UMLS supplemented with automatically translated English UMLS terms to index the QUAERO corpus. Several post-processing steps were implemented to reduce the number of false positive detections, including filtering based on precision scores that were derived from the training data. For the coding task, two ICD-10 terminologies were constructed based on the training material that was supplied by the challenge organizers. The Solr text tagger was used with these terminologies to index the death certificates and generate codes. Again, precision-score filtering was applied to improve precision.

LITL The LITL team participated in the plain entity recognition task [25]. The LITL team system was specifically designed by master's students (LITL programme, university of Toulouse) and their teachers for the challenge. The system used is mainly based on supervised machine learning, through the use of a CRF classifier (CRF++ 0.58) based on a varierty of linguistic features (Part-Of-Speech tags, generic word lists and syntactic parsing). Training and test data have been POS-tagged and parsed by the Talismane toolkit [26], and external resources were used to tag the tokens (generic lists of suffixes and prefixes, word lists from SNOMED and from VIDAL database). The output of the CRF was completed by a custom-made rule-based system which identifies syntactic patterns in order to extract more complex entities.

LIMSI The LIMSI team participated in the coding task [27]. Their system offered a classifier with humanly-interpretable output, based on IR-style ranking of candidate ICD10 diagnoses. A tf.idf-weighted bag-of-feature vector was built for each training set code by merging all the statements found for this code in the training data. Given a new statement, candidate codes were ranked with Cosine similarity. Features included meta-information and n-grams of normalized tokens. An ICD chapter classifier was also prepared with the same method and it was used to rerank the top-k codes (k=2) returned by the code classifier. The development phase focused on mono-code statements. Good precision could be obtained using the top code and a significant performance gain was yielded by chapter reranking. Accordingly, on test data, the system was set to return one code for each statement, leaving multiple code assignment for future work.

SIBM The SIBM team participated in all tasks [28]. They approached entity extraction from the provided QUAERO dataset as an indexing task relying on multiple knowledge organization systems (KOS) partially or totally translated into French. The extraction method, ECMT (Extracting Concepts with Multiple Terminologies), performs bag of words concept matching at the sentence level. It was originally designed to extract clinical concepts from Electronic Health Records. They addressed the identification of relevant clinical entities within the International Classification of Diseases version 10 in the CépiDC dataset with the CIMIND system based on natural language processing and approximate string matching methods.

UPF The UPF team participated in the plain entity recognition and the normalization tasks [29]. They proposed two different systems for solving each phase. For Phase I (entity recognition), a basic system uses a distant learning approach based on a set SVM classifiers (one for each class) followed by a voting scheme for choosing the best result. A second run was also submitted combining the result of the basic system (run 1) with some symbolic processing for improving entity classification. In Phase II (entity normalization), the system obtains normalization information from public resources after obtaining the English translation of each medical term.

3.2 System performance on entity recognition

Tables 4 and 5 present system performance on the plain entity recognition task. Tables 6 and 7 present system performance on the normalized entity recognition task. Team Erasmus had the best performance in terms of F-measure for both the EMEA and MEDLINE corpora with their official runs. However, an unofficial run (shown in italic font) submitted after the challenge deadline outperformed the official runs by using the Solr indexing method instead of Peregrine. This suggests that for knowledge-based methods, the specific method used for matching lexical resources carries a significant weight, in addition to the coverage of these resources. Team LITL reports performing some corrective pre-processing of the text to address extraneous spaces ocurring around punctuation marks, which may cause issues with entity or concept recognition. However, they do not report on the impact of the corrective step on their system performance. Compared to last year, this year's performance show that all technical difficulties linked to the corpus format and annotation format seem to have been resolved.

A t-test comparing all pairs of runs at entity level showed that all differences between runs were significant ($p < 0.001$), with the exception of the two runs from LILT ($p = 0.28$ on EMEA, exact match, $p = 0.73$ on MEDLINE, exact match).

Table 4. System performance for plain entity recognition on the EMEA test corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-measure**
*Erasmus-run3.unofficial**	*1729*	*685*	*475*	*0.716*	*0.785*	***0.749***
Erasmus-run2*	1732	1001	472	0.634	0.786	0.702
Erasmus-run1*	1757	1063	447	0.623	**0.797**	0.699
LITL-run1*	879	242	1325	**0.784**	0.399	0.529
LITL-run2	867	264	1337	0.767	0.393	0.520
SIBM-run1*	834	716	1370	0.538	0.378	0.444
SIBM-run2*	724	483	1480	0.600	0.329	0.425
BITEM-run1*	406	371	1798	0.523	0,184	0.272
UPF-run1*	512	3463	1835	0.129	0.218	0.162
*UPF-run2.unofficial**	*420*	*4025*	*1816*	*0.095*	*0.188*	*0.126*
average				0.575	0.436	0.469
median				0.611	0.386	0.482

3.3 System performance on entity normalization

Tables 8 and 9 present system performance on the entity normalization task. Team SIBM had the best performance in terms of F-measure for both the EMEA

Table 5. System performance for plain entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
*Erasmus-run3.unofficial**	*2220*	*1045*	*881*	***0.680***	***0.716***	***0.698***
Erasmus-run1*	2139	1330	962	0.617	0.690	0.651
Erasmus-run2*	2103	1273	998	0.623	0.678	0.649
SIBM-run2*	1357	761	1745	0.641	0.438	0.520
SIBM-run1*	1476	1258	1626	0.540	0.476	0.506
BITEM-run1*	1376	1032	1741	0.571	0.442	0.498
LITL-run1*	998	556	2105	0.642	0.322	0.429
LITL-run2	989	561	2114	0.638	0.319	0.425
*UPF-run2.unofficial**	*969*	*5050*	*2138*	*0.161*	*0.312*	*0.212*
UPF-run1*	736	5053	2369	0.127	0.237	0.166
UPF-run2	739	5050	2367	0.128	0.238	0.166
average				0.503	0.426	0.446
median				0.617	0.438	0.498

Table 6. System performance for normalized entity recognition on the EMEA test corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
*Erasmus-run3.unofficial**	*1542*	*672*	*872*	***0.697***	***0.639***	***0.666***
Erasmus-run1*	1630	1709	1190	0.488	0.578	0.529
Erasmus-run2*	1607	1732	1126	0.481	0.588	0.529
SIBM-run1*	592	1611	966	0.269	0.380	0.315
SIBM-run2*	467	1736	735	0.212	0.389	0.274
BITEM-run1*	347	1856	430	0.158	0.447	0.233
average				0.322	0.476	0.376
median				0.269	0.477	0.315

Table 7. System performance for normalized entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents runs that were submitted after the official deadline. The median and average are computed solely using the official runs. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
*Erasmus-run3.unofficial**	*1943*	*1320*	*1220*	***0.596***	***0.614***	***0.605***
Erasmus-run1*	1948	2802	1519	0.410	0.562	0.474
Erasmus-run2*	1917	2833	1457	0.404	0.568	0.472
BITEM-run1*	1187	1911	1220	0.383	0.4931	0.431
SIBM-run2*	1012	2083	1108	0.327	0.477	0.388
SIBM-run1*	1102	1993	1638	0.356	0.402	0.378
average				0.376	0.501	0.429
median				0.383	0.493	0.431

and MEDLINE corpora, using a combination of knowledges resources dedicated to French, compared to team UPF which relied on matching a translation of the terms into English to English resources.

Table 8. System performance for entity normalization on the EMEA test corpus. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
SIBM-run2*	1019	667	1184	**0.604**	0.463	**0.524**
SIBM-run1*	1047	800	1156	0.567	**0.475**	0.517
UPF-run1*	517	558	558	0.481	0.481	0.481
average				0.551	0.473	0.507
median				0.567	0.475	0.517

3.4 System performance on death certificate coding

Table 10 presents system performance on the ICD10 coding task. Team Erasmus had the best performance in terms of F-measure. Overall, systems performed high on the coding task. It is interesting to note that participants addressed this task independently from the entity recognition and normalization tasks offered on the QUAERO corpus. Since ICD10 is one of the terminologies aggregated within the UMLS, a reasonable approach might have been to extract UMLS concepts from the text of death certificates, and then restrict the results to ICD10 in order to produce coding recommendations. However, none of the participating teams chose this approach. The results show that both knowledge-based and statistical methods can perform well on the task, as the best performance is

Table 9. System performance for entity normalization on the MEDLINE test corpus. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
SIBM-run1*	1598	1094	1505	0.594	**0.515**	**0.552**
SIBM-run2*	1450	978	1651	**0.597**	0.468	0.524
UPF-run1*	673	745	748	0.475	0.474	0.474
average				0.555	0.485	0.568
median				0.594	0.474	0.525

obtained from a knowledge-based method, while the second best is obtained with statistical methods (Team ERIC-ECSTRA), followed by another knowledge based method (team SIBM). The results are very encouraging from a practical perspective and indicate that a coding assistance system could prove very useful for the effective processing of death certificates.

Table 10. System performance for ICD10 coding on the CépiDC test corpus. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, per student test.

Team	TP	FP	FN	Precision	Recall	F-measure
Erasmus-run2*	88497	11423	20321	0.886	**0.813**	**0.848**
Erasmus-run1*	87404	10823	21414	**0.890**	0.803	0.844
ERIC-ECSTRA-run2*	71319	9479	37499	0.882	0.655	0.752
ERIC-ECSTRA-run1*	66954	15605	41864	0.811	0.615	0.700
SIBM-run1*	72192	31480	36626	0.696	0.663	0.680
LIMSI-run1*	61874	19002	46984	0.765	0.569	0.652
BITEM-run1*	57256	40650	51562	0.585	0.526	0.554
average				0.788	0.664	0.719
median				0.811	0.655	0.700

3.5 Replication track and replicability of the results

Three teams submitted systems to our replication track: one system covered both QUAERO and CépiDC data, and two systems only processed CépiDC data. Two teams expressed interest in submitting a system but eventually reported that they did not have time to make the system ready for submission. One team reported that they were reserving the distribution of their system to commercial use and one team did not provide a reason for not participating to the track.

The system submitted for replicating QUAERO results was in fact incomplete as the submission included the results of pre-processing the corpus with a tool that the team did not share as part of the replication track. Between the

two analysts working with each system, we were able to replicate exactly the results submitted by 6 of the target runs (the QUAERO runs and two CépiDC runs): the precision, recall and F-measure obtained from running the systems were identical to that of the runs submitted by participants. For one run adressing the CépiDC corpus, only one analyst was able obtain results from the system, and the results obtained showed a 0.02 difference in F-measure, which was statistically significant. The analysts experienced varying degrees of difficulty to install and run the systems. Differences were mainly due to the technical set-up of the computers used to replicate the experiments. Analysts also report that additional information on system requirements, installation procedure and practical use would be useful for all the systems submitted. Overall, this indicates that replication is achievable. However, it is not as straight-forward as one would hope. More detailed communication about the systems could be an important step towards making replication an effortless reality.

4 Conclusion

We released a new portion of the QUAERO French Medical corpus through Task 2 of the CLEFeHealth 2016 Evaluation Lab. This corpus contains entity annotations for ten entities of clinical interest, with normalization to UMLS CUIs. In the evaluation lab, we evaluated systems on the task of plain or normalized entity recognition as well as on the task of assigning CUIs to pre-identified entities (normalization). In addition, we also released a large corpus of French death certificates to evaluate systems on the task of ICD10 coding. This is the second edition of a biomedical NLP challenge that provides large gold-standard annotated corpora in French. Results show that high performance can be achieved by NLP systems on the tasks of entity recognition, normalization and coding for French biomedical text. The corpus used and the participating team system results are an important contribution to the research community and the focus on a language other than English (French) remains a rare initiative.

Acknowledgements

We want to thank all participating teams for their effort in addressing new and challenging tasks. We also want to thank Jan Kors from team Erasmus for his contribution to the CépiDC evaluation script. The organization work for CLEF eHealth 2016 task 2 was supported by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-13-JCJC-SIMI2-CABeRneT.

The CLEF eHealth 2016 evaluation lab has been supported in part by (in alphabetical order) PhysioNetWorks Workspaces; the CLEF Initiative;

References

1. Jones KS, Galliers JR. Evaluating natural language processing systems: An analysis and review. 1995. Springer Science & Business Media:1083

2. Voorhees EM, Harman DK and others. TREC: Experiment and evaluation in information retrieval, vol 1. 2005. MIT press Cambridge.
3. Suominen H, Salantera S, Velupillai S, Chapman WW, Savova G, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJF, Leveling J, Kelly L, Goeuriot L, Martinez D, Zuccon G. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds), Information Access Evaluation. Multilinguality, Multimodality, and Visualization. LNCS (vol. 8138):212-231. Springer, 2013
4. 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: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer, 2015
5. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, Velupillai S, Chapman WW, Martinez D, Zuccon G, Palotti J. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds), Information Access Evaluation. Multilinguality, Multimodality, and Interaction. LNCS (vol. 8685):172-191. Springer, 2014
6. Névéol A, Grouin C, Tannier X, Hamon T, Kelly L, Goeuriot L, Zweigenbaum P (2015). CLEF eHealth Evaluation Lab 2015 Task 1b: clinical named entity recognition. CLEF 2015, Online Working Notes, CEUR-WS 1391.
7. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc, 18(5):540-3
8. Huang CC, Lu Z (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform, 2015 May 1. pii: bbv024.
9. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics. 2005;6 Suppl 1:S2.
10. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, Lin YS, Klinger R, Friedrich CM, Ganchev K, Torii M, Liu H, Haddow B, Struble CA, Povinelli RJ, Vlachos A, Baumgartner WA Jr, Hunter L, Carpenter B, Tsai RT, Dai HJ, Liu F, Chen Y, Sun C, Katrenko S, Adriaans P, Blaschke C, Torres R, Neves M, Nakov P, Divoli A, Maña-López M, Mata J, Wilbur WJ. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;9 Suppl 2:S2
11. Arighi CN, Wu CH, Cohen KB, Hirschman L, Krallinger M, Valencia A, Lu Z, Wilbur JW, Wiegers TC. BioCreative-IV virtual issue. Database (Oxford). 2014 May 22;2014. pii: bau039.
12. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):552-6.
13. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics. 2005;6 Suppl 1:S11.
14. , Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu HH, Torres R, Krauthammer M, Lau WW, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L. Overview of BioCreative II gene normalization. Genome Biol. 2008;9 Suppl 2:S3.
15. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, Hsu CN, Tsai RT, Dai HJ, Okazaki N, Cho HC, Gerner M, Solt I, Agarwal S, Liu F, Vishnyakova D, Ruch P, Romacker M, Rinaldi F, Bhattacharya S, Srinivasan P, Liu H, Torii M, Matos S, Campos D, Verspoor K, Livingston KM, Wilbur WJ. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011 Oct 3;12 Suppl 8:S2.

16. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):552-6.
17. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P (2014). Language Resources for French in the Biomedical Domain. In: Proc of LREC, p. 2146-2151
18. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P (2014). The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In: Proc of Bio TextM, p. 24-30
19. Pavillon G., Laurent F (2003). Certification et codification des causes médicales de décés. Bulletin Epidémiologique Hebdomadaire - BEH:134-138. `http://opac.invs.sante.fr/doc_num.php?explnum_id=2065` (accessed: 2016-06-06)
20. Verspoor K, Jimeno Yepes A, Cavedon L, McIntosh T, Herten-Crabb H, Thomas Z, Plazzer JP (2013). Annotating the Biomedical Literature for the Human Variome. Database (Oxford), virtual issue for BioCuration 2013 meeting
21. Mottin L, Gobeill J, Mottaz A, Pasche E, Gaudinat A, Ruch P (2016). BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction CLEF 2016 Online Working Notes. CEUR-WS
22. Dermouche M, Looten V, Flicoteaux R, Chevret S, Velcin J and Taright N (2016). ECSTRA-INSERM @ CLEF eHealth2016-task 2: ICD10 Code Extraction from Death Certificates. CLEF 2016 Online Working Notes. CEUR-WS
23. Van Mulligen E, Afzal Z, Akhondi SA, Vo D, Kors JA (2016). Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts. CLEF 2016 Online Working Notes, CEUR-WS
24. Afzal Z, Akhondi SA, van Haagen H, Van Mulligen E and Kors JA (2015). Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. CLEF 2015 Online Working Notes. CEUR-WS
25. Ho-Dac LM, Tanguy L, Grauby C, Hnub N, Heu Mby A, Malosse J, Rivière L, Veltz-Mauclair A and Wauquier M (2015). LITL at CLEF eHealth2016: recognizing entities in French biomedical documents. CLEF 2016 Online Working Notes. CEUR-WS
26. Urieli A (2013). Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit. PhD thesis. Université de Toulouse II-Le Mirail
27. Zweigenbaum P and Lavergne T (2016). LIMSI ICD10 coding experiments on CépiDC death certificate statements. CLEF 2016 Online Working Notes. CEUR-WS
28. Cabot C, Soualmia LF, Dahamna B and Darmoni SJ (2016). SIBM at CLEF eHealth Evaluation Lab 2016: Extracting Concepts in French Medical Texts with ECMT and CIMIND. CLEF 2016 Online Working Notes. CEUR-WS
29. Vivaldi J, Rodriguez H and Cotik V (2016). Semantic tagging and normalization of French medical entities. CLEF 2016 Online Working Notes. CEUR-WS