# CLEF eHealth Evaluation Lab 2015 Task 1b: clinical named entity recognition

Aurélie Névéol[1], Cyril Grouin[1], Xavier Tannier[2,1], Thierry Hamon[1,3]
Liadh Kelly[4], Lorraine Goeuriot[5], and Pierre Zweigenbaum[1]

[1] LIMSI–CNRS, UPR 3251, Orsay, France
firstname.lastname@limsi.fr
[2] Université Paris-Sud, Orsay, France
[3] Université Paris Nord, Villetaneuse, France
[4] ADAPT Centre, Trinity College, Dublin, Ireland
liadh.kelly@tcd.ie
[5] Université Grenoble Alpes, Grenoble, France
lorraine.goeuriot@imag.fr

**Abstract.** This paper reports on Task 1b of the 2015 CLEF eHealth evaluation lab which extended the previous information extraction tasks of ShARe/CLEF eHealth evaluation labs by considering ten types of entities including *disorders*, that were to be extracted from biomedical text in French. The task consisted of two phases: entity recognition (phase 1), in which participants could supply plain or normalized entities, and entity normalization (phase 2). The entities to be extracted were defined according to Semantic Groups in the Unified Medical Language System® (UMLS®), which was also used for normalizing the entities. Participant systems were evaluated against a blind reference standard of 832 titles of scientific articles indexed in MEDLINE and 3 full text drug monographs published by the European Medicines Agency (EMEA) using Precision, Recall and F-measure. In total, seven teams participated in phase 1, and three teams in phase 2. The highest performance was obtained on the EMEA corpus, with an overall F-measure of 0.756 for plain entity recognition, 0.711 for normalized entity recognition and 0.872 for entity normalization.

**Keywords:** Natural Language Processing; Information Extraction; Named Entity Recognition, Concept Normalization, UMLS, French, Biomedical Text

## 1 Introduction

Following healthcare laws that grant patients access to their own medical information in the United States [2] and in Europe [1], the information extraction (IE) challenges in the CLEF eHealth Lab have strived to put forth tools and methods to help patients understand the content of their health records. Over the past two years, these IE challenges have addressed named entity recognition, normalization [3] and attribute extraction [4]. The focus was on a widely studied

type of corpus, namely written English clinical text [3, 4]. This year the lab's [5] IE challenge evolved to address lesser studied corpora, including biomedical texts in a language other than English. Languages other than English were previously featured in a multilingual context in the recent CLEF-ER 2013 lab [6]. However, the task offered in CLEF eHealth 2015 Task 1b is the first shared task based on a large gold standard annotated biomedical corpus in a language other than English.

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, and entity normalization as evidenced by the previous CLEF eHealth labs [3, 4]. 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 in the biomedical domain, compared to English [9].

## 2 Material and Methods

We describe the dataset, the tasks and the evaluation metrics used for the CLEF eHealth 2015 Evaluation Lab Task 1b.

### 2.1 Dataset

**Description of the annotated data** The data set is called QUAERO French Medical Corpus. It was developed as a resource for named entity recognition and normalization in 2013 [10].

The data set was created in the wake of the 2013 CLEF-ER challenge [6], 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 submitted for human annotation. The annotation process was guided by concepts in the Unified Medical Language System (UMLS):

- 10 types of clinical entities, as defined by the following UMLS Semantic Groups [11], were annotated: Anatomy, Chemicals & 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: (a) 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); (b) 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"); (c) An entity whose 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).

Significant work was done on the initial QUAERO French Medical Corpus in order to convert the annotation format from an in-line XML format to a stand-off format relying on text character offsets. In the process, annotation errors were corrected, which included systematic checking of annotation format and the introduction of discontinuous entity annotations. While discontinuous entities were part of the original QUAERO French Medical Corpus annotation guidelines, they had been poorly marked due to technical difficulties linked to the in-line XML format.

Annotations on the training set are provided in the BRAT standoff format [12] and can be visualized using the BRAT Rapid Annotation Tool [13]. Participants were also expected to supply annotations in this format.

The training set released in the CLEF eHealth 2015 Task 1b challenge comprised 833 MEDLINE titles and 3 full EMEA documents (divided between 11 files for readability through the BRAT interface). The test set comprised 832 MEDLINE titles and 3 full EMEA documents (divided into 12 files). Table 1 presents additional statistics describing the corpus contents.

**Table 1.** Descriptive statistics of the corpus

| | EMEA | | MEDLINE | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Tokens | 14,944 | 13,271 | 10,552 | 10,503 |
| Entities | 2,695 | 2,260 | 2,994 | 2,977 |
| Unique Entities | 923 | 756 | 2,296 | 2,288 |
| Unique CUIs | 648 | 523 | 1,860 | 1,848 |

**Dataset Excerpts** Figure 1 shows two sample MEDLINE documents with the corresponding complete annotations for normalized entities. Figure 2 shows an excerpt of one EMEA document with the corresponding complete annotations for normalized entities.

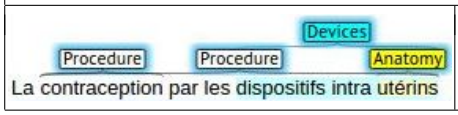
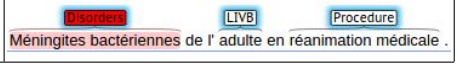
| MEDLINE title 1 | MEDLINE title 2 |
|---|---|
| *La contraception par les dispositifs intra utérins* .[1] | *Méningites bactériennes de l' adulte en réanimation médicale* .[2] |
| MEDLINE title 1 annotations | MEDLINE title 2 annotations |
| T1 PROC 3 16 contraception<br>#1 AnnotatorNotes T1 C0700589<br>T2 DEVI 25 50 dispositifs intra utérins<br>#2 AnnotatorNotes T2 C0021900<br>T3 ANAT 43 50 utérins<br>#3 AnnotatorNotes T3 C0042149 | T1 DISO 0 23 Méningites bactériennes<br>#1 AnnotatorNotes T1 C0085437<br>T2 LIVB 29 36 adulte<br>#2 AnnotatorNotes T2 C0001765<br>T3 PROC 40 60 réanimation médicale<br>#3 AnnotatorNotes T3 C0085559 |
| Document view using BRAT | |
|  |  |

**Fig. 1.** Sample annotated MEDLINE titles

### 2.2 Tasks

**Named entity recognition.** 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 (see Section 2.1). Participants could mark either plain entities (i.e. 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.** The task of entity normalization consisted of mapping entities of clinical interest marked in biomedical text to a relevant UMLS CUI.

### 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 $\beta$=1.) for named entity recognition and entity normalization.

[1] Contraception by intrauterine devices
[2] Bacterial meningitis in adults in the intensive care unit.
[3] What is Tysabri used for? Tysabri is used to treat adults with highly active multiple sclerosis (MS).

EMEA document (excerpt)

(...)
*Dans quel cas Tysabri est-il utilisé ?*
*Tysabri est utilisé dans le traitement des adultes atteints de sclérose en plaques ( SEP ).*[3]
(...)

EMEA document annotations (excerpt)

(...)
T9 CHEM 206 213 Tysabri
#9 AnnotatorNotes T9 C1529600
T10 CHEM 233 240 Tysabri
#10 AnnotatorNotes T10 C1529600
T11 PROC 261 271 traitement
#11 AnnotatorNotes T11 C0087111
T12 LIVB 276 283 adultes
#12 AnnotatorNotes T12 C0001675
T13 DISO 296 315 sclérose en plaques
#13 AnnotatorNotes T13 C0026769
T14 DISO 318 321 SEP
#14 AnnotatorNotes T14 C0026769
(...)

Document view using BRAT


**Fig. 2.** Excerpt of a sample annotated EMEA document

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

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

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

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. Results were computed using the brateval program initially developed by Verspoor et al. [14], which we extended to cover the evaluation of normalized entities. The updated version of brateval was 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.

For **normalized entity recognition**, an exact match (true positive) was counted when the system's entity type, span and CUIs matched the reference.

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.

## 3 Results

Participating teams included between one and six team members and resided in Belarus (team IHS-RD), China (team HIT-WI), France (teams CISMeF and LIMSI), India (team Watchdogs), the Netherlands (team Erasmus) and Spain (Team UPF).

For the plain entity recognition task, seven teams submitted a total of 10 runs for each of the corpora, EMEA and MEDLINE (20 runs in total). For the normalized entity recognition task, four teams submitted a total of 5 runs for each of the corpora (10 runs in total). For the normalization task, three teams submitted a total of 4 runs for each of the corpora (8 runs in total).

### 3.1 Methods implemented in the participants' systems

Participants used a variety of methods, some of which used machine-learning. Non-machine learning methods relied on lexical sources (medical terminologies and ontologies), translation software (statistical machine translation) or a combination of both and did not use the training corpus at all. Machine-learning methods relied on Conditional Random Fields (CRFs) for entity recognition, and used lexical resources as features. Many participants used the standard distribution of the UMLS as an onto-terminological resource. Teams from France used additional UMLS-related resources for French. Other teams relied on other sources such as Wikipedia, MANTRA and term translations provided by statistical machine translation tools. It can be noted that for the entity recognition task, none of the participants sought to address the case of discontinuous entities.

The CISMeF team participated in the plain and normalized entity recognition subtasks[15]. They trained CRF models for each entity type and each of the two corpora. They used lexical, part-of-speech and orthographical features, including 4-character prefixes and suffixes for the current word and neighboring words. A lexicon of geographical names was used independently of the CRF to tag geographical entities. UMLS information was used only to identify CUIs, not to produce features.

The Erasmus team participated in all three subtasks [16]. They trained CRF models for each entity type and each of the two corpora. They used lexical, part-of-speech and orthographical features, including 4-character prefixes and suffixes for the current word and neighboring words. A lexicon of geographical names was used independently of the CRF to tag geographical entities. UMLS information was used only to identify CUIs, not to produce features.

The HIT-WI team participated in all three subtasks [17]. They trained CRF models for each entity type and each of the two corpora. They used lexical, part-of-speech and orthographical features, including 4-character prefixes and suffixes for the current word and neighboring words. A lexicon of geographical names was used independently of the CRF to tag geographical entities. UMLS information was used only to identify CUIs, not to produce features.

The IHS-RD team participated in all three subtasks [18]; however, they focused their efforts on the plain entity recognition subtask. They authors built 10 binary classifiers with the same sets of features: uni-grams and bi-grams and associated information such as case of the strings, presence of non-alphabetic characters, part-of-speech, syntactic function, UMLS semantic categories and occurrence in the general language. Their analysis of the contribution of the type of features shows that UMLS semantic categories have a strong impact on the results, while the contribution of syntactic features depends on the corpus.

The LIMSI team participated in the plain entity recognition subtask [19]. LIMSI's identification system is based on the combination of three classifiers, in order to deal with embedded entities (16% of entities in the training set): a first CRF detects non-embedded entities, a second context-free CRF detects embedded entities, and a SVM identifies their semantic class. These classifiers rely on a set of features used in state-of-the-art classification systems, including token/POS ngrams, morphologic features, and dictionary consultation in language-dependent external sources.

The UPF team participated in the plain entity recognition subtask [20]. The team used an existing system, designed to annotate medical entities in English, based on a distant learning approach. Their goal was to evaluate the robustness of their method on a corps in language other than English, and the system was used "out of box", without using the training data. The method relies on several SVM classifiers (one per category) and a voting procedure (the best score) to select the result from all classifiers. Classifiers are not trained on the training corpus but on resources produced from external resources (French Wikipedia).

The Watchdog team participated in the plain entity recognition subtask [21]. Their system (Run 1) used a CRF on stemmed tokens with standard lexical features and the word position in the sentence (discretized into three bins). The originality of the approach is that UMLS features were obtained by translating words from French to English with the Bing translator before applying MetaMap on the resulting English words to obtain their semantic groups. A variant of the system (Run 2) directly used the UMLS features to detect entities for words where the CRF detected no entity, but performs less well than the initial method (Run 1).

### 3.2 System performance on entity recognition

Tables 2 and 3 present system performance on the plain entity recognition task. Tables 4 and 5 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. An analysis of the results showed that

entity offsets were a technical difficulty for many teams and resulted in zero or close-to-zero performance for runs exhibited formatting issues. To gain a better insight of method performance, we invited participants to submitted revised versions of their runs, where offset and formatting issues had been corrected. These submissions occurred after the lab deadline, and are shown in italic font in the tables.

Overall, in official runs, systems performed higher on the MEDLINE corpus (average F-measure of 0.396 for plain entities, and 0.336 for normalized entities) compared to EMEA (average F-measure of 0.279 for plain entities, and 0.311 for normalized entities). However, once format fixes are taken into account, system performance is in fact higher on EMEA documents. This is explained by the fact that MEDLINE titles were short documents, comprising only one or two sentences at most. EMEA documents were much longer (several hundred sentences) and offset errors in some official runs often occurred beyond the first sentence. Once the formatting issues are corrected, it appears that systems perform better on EMEA documents, which are much more redundant than MEDLINE titles.

**Table 2.** System performance for plain entity recognition on the EMEA test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs.

| **Team** | **TP** | **FP** | **FN** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|---|---|---|
| Erasmus-run1 | 1720 | 570 | 540 | 0.751 | 0.761 | **0.756** |
| Erasmus-run2 | 1753 | 716 | 507 | 0.710 | **0.776** | 0.741 |
| *IHS-RD-run1-fix* | *1350* | *223* | *910* | *0.858* | *0.597* | *0.704* |
| Watchdogs-run1 | 1238 | 203 | 1022 | **0.859** | 0.548 | 0.669 |
| *IHS-RD-run2-fix* | *1288* | *328* | *972* | *0.797* | *0.570* | *0.665* |
| *HIT-WI Lab-run1-fix* | *971* | *234* | *1289* | *0.806* | *0.430* | *0.561* |
| LIMSI-run1 | 945 | 644 | 1315 | 0.595 | 0.418 | 0.491 |
| Watchdogs-run2 | 1309 | 2361 | 951 | 0.357 | 0.579 | 0.442 |
| *UPF-run1-fix* | *113* | *2147* | *704* | *0,050* | *0,138* | *0,073* |
| HIT-WI Lab-run1 | 12 | 1137 | 2248 | 0.010 | 0.005 | 0.007 |
| CISMeF-run1 | 9 | 4124 | 2251 | 0.002 | 0.004 | 0.003 |
| IHS-RD-run1 | 0 | 0 | 2260 | 0.000 | 0.000 | 0.000 |
| IHS-RD-run2 | 0 | 1616 | 2260 | 0.000 | 0.000 | 0.000 |
| UPF-run1 | 0 | 1067 | 2260 | 0.000 | 0.000 | 0.000 |
| **average (official)** | | | | 0.328 | 0.309 | 0.311 |
| *average-fix* | | | | 0.587 | 0.473 | 0.511 |
| **median (official)** | | | | 0.184 | 0.212 | 0.224 |
| *median-fix* | | | | 0.731 | 0.559 | 0.613 |

**Table 3.** System performance for plain entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs.

| **Team** | **TP** | **FP** | **FN** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|---|---|---|
| Erasmus-run1 | 1861 | 756 | 1116 | 0.711 | 0.625 | **0.665** |
| Erasmus-run2 | 1912 | 886 | 1065 | 0.683 | **0.642** | 0.662 |
| *IHS-RD-run1-fix* | *1195* | *1782* | *376* | *0.761* | *0.401* | *0.526* |
| IHS-RD-run2 | 1188 | 383 | 1789 | **0.756** | 0.399 | 0.522 |
| Watchdogs-run1 | 1215 | 490 | 1762 | 0.713 | 0.408 | 0.519 |
| LIMSI-run1 | 1121 | 834 | 1856 | 0.573 | 0.377 | 0.455 |
| HIT-WI Lab-run1 | 1068 | 671 | 1909 | 0.614 | 0.359 | 0.453 |
| Watchdogs-run2 | 1364 | 2069 | 1613 | 0.397 | 0.458 | 0.426 |
| CISMeF-run1 | 680 | 4412 | 2297 | 0.134 | 0.228 | 0.169 |
| *UPF-run1-fix* | *189* | *2788* | *817* | *0,064* | *0,188* | *0,095* |
| IHS-RD-run1 | 75 | 168 | 2902 | 0.309 | 0.025 | 0.047 |
| UPF-run1 | 82 | 888 | 2895 | 0.085 | 0.028 | 0.042 |
| **average (official)** | | | | 0.498 | 0.355 | 0.396 |
| *average-fix* | | | | 0.553 | 0.396 | 0.440 |
| **median (official)** | | | | 0.594 | 0.388 | 0.454 |
| *median-fix* | | | | 0.649 | 0.400 | 0.487 |

**Table 4.** System performance for normalized entity recognition on the EMEA test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs.

| **Team** | **TP** | **FP** | **FN** | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|---|---|---|
| CISMeF-run1 | 10 | 2255 | 4128 | 0.004 | 0.002 | 0.003 |
| Erasmus-run1 | 1637 | 655 | 678 | **0.714** | **0.707** | **0.711** |
| Erasmus-run2 | 1627 | 680 | 866 | 0.705 | 0.653 | 0.678 |
| IHS-RD-run1 | 0 | 2260 | 1616 | 0.000 | 0.000 | 0.000 |
| *IHS-RD-run1-fix* | *923* | *17264* | *710* | *0.051* | *0.565* | *0.093* |
| HIT-WI Lab-run1 | 8 | 2252 | 1112 | 0.003 | 0.007 | 0.005 |
| *HIT-WI Lab-run1-fix* | *432* | *1828* | *735* | *0.191* | *0.370* | *0,252* |
| **average (official)** | | | | 0.286 | 0.274 | 0.279 |
| *average-fix* | | | | 0.333 | 0.460 | 0.347 |
| **median (official)** | | | | 0.004 | 0.007 | 0.005 |
| *median-fix* | | | | 0.191 | 0.565 | 0.252 |

**Table 5.** System performance for normalized entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs.

| Team | TP | FP | FN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| CISMeF-run1 | 1020 | 2434 | 4461 | 0.295 | 0.186 | 0.228 |
| Erasmus-run1 | 1660 | 1376 | 957 | **0.547** | **0.634** | **0.587** |
| Erasmus-run2 | 1677 | 1363 | 1121 | 0.552 | 0.599 | 0.575 |
| IHS-RD-run1 | 634 | 15170 | 938 | 0.040 | 0.403 | 0.073 |
| *IHS-RD-run1-fix* | *927* | *17495* | *644* | *0.050* | *0.590* | *0.093* |
| HIT-WI Lab-run1 | 515 | 2460 | 1223 | 0.173 | 0.2963 | 0.219 |
| **average (official)** | | | | 0.321 | 0.424 | 0.336 |
| *average-fix* | | | | 0.323 | 0.461 | 0.340 |
| **median (official)** | | | | 0.295 | 0.403 | 0.228 |
| *median-fix* | | | | 0.295 | 0.590 | 0.228 |

### 3.3 System performance on entity normalization

Tables 6 and 7 present system performance on the entity normalization task. Team Erasmus had the best performance in terms of F-measure for both the EMEA and MEDLINE corpora. Overall, systems performed higher on the EMEA corpus (average F-measure of 0.615) compared to MEDLINE (average F-measure of 0.475). It can be explained by the fact that entities in the EMEA corpus are much more redundant compared to the MEDLINE corpus (see Unique CUIs counts in Table 1).

**Table 6.** System performance for entity normalization on the EMEA test corpus

| Team | TP | FP | FN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| Erasmus-run1 | 1734 | 526 | 0 | 0.767 | **1.000** | 0.868 |
| Erasmus-run2 | 1748 | 512 | 0 | **0.774** | **1.000** | **0.872** |
| IHS-RD-run1 | 1578 | 26642 | 715 | 0.056 | 0.688 | 0.103 |
| HIT-WI Lab-run1 | 1266 | 994 | 1027 | 0.560 | 0.552 | 0.556 |
| **average (official)** | | | | 0.532 | 0.896 | 0.615 |
| **median (official)** | | | | 0.767 | 1.000 | 0.868 |

## 4 Discussion and Conclusion

We released an improved version of the QUAERO French Medical corpus through Task 1b of the CLEFeHealth 2015 Evaluation Lab. This corpus contains entity annotations for ten entities of clinical interest, with normalization to UMLS

**Table 7.** System performance for entity normalization on the MEDLINE test corpus

| Team | TP | FP | FN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| Erasmus-run1 | 1780 | 1328 | 398 | 0.573 | **0.817** | **0.674** |
| Erasmus-run2 | 1787 | 1321 | 433 | **0.575** | 0.805 | 0.671 |
| IHS-RD-run1 | 1712 | 38213 | 1264 | 0.043 | 0.575 | 0.080 |
| HIT-WI Lab-run1 | 1386 | 1589 | 1590 | 0.466 | 0.466 | 0.466 |
| **average (official)** | | | | 0.397 | 0.733 | 0.475 |
| **median (official)** | | | | 0.573 | 0.805 | 0.671 |

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). This is a unique biomedical NLP challenge—no previous challenge has provided such a large gold-standard annotated corpus in a language other than English. Results show that high performance can be achieved by NLP systems on the task of entity recognition and normalization for French biomedical text. However performance levels varied greatly between participating teams, indicating that the tasks are highly challenging. This corpus and the participating team system results are an important contribution to the research community and the focus on a language other than English (French) is unprecedented.

## Acknowledgements

We want to thank all participating teams for their effort in addressing a new and challenging task. We also want to thank Afzal Zubair from team Erasmus for his extensive testing of the evaluation script. The organization work for CLEF eHealth 2015 task 1B was supported by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-13-JCJC-SIMI2-CABeRneT.

The CLEF eHealth 2015 evaluation lab has been supported in part by (in alphabetical order) PhysioNetWorks Workspaces; the CLEF Initiative;

## References

1. European Union: Directive 2011/24/EU of the European Parliament and of the Council of 9 march 2011. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:088:0045:0065:en:PDF` Accessed: 2015-05-13.
2. Center for Medicare, Medicaid Services: Eligible professional meaningful use menu set measures: Measure 5 of 10. `http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/5_Patient_Electronic_Access.pdf` Accessed: 2015-05-13.
3. Hanna Suominen, Sanna Salantera, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot,

David Martinez, Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Pamela Forner, Henning Muller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation. Multilinguality, Multimodality, and Visualization, volume 8138 of Lecture Notes in Computer Science, pages 212– 231. Springer Berlin Heidelberg, 2013.

4. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martinez, Guido Zuccon, João Palotti Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury and Elaine Toms, editors, Information Access Evaluation. Multilinguality, Multimodality, and Interaction, volume 8685 of Lecture Notes in Computer Science, pages 172-191. Springer International Publishing, 2014.
5. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, Joao Palotti, Guido Zuccon Overview of the CLEF eHealth Evaluation Lab 2015. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer International Publishing, 2015.
6. Dietrich Rebholz-Schuhmann, Simon Clematide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and Jan A. Kors. Entity recognition in parallel multilingual biomedical corpora : The CLEF-ER laboratory overview. In Pamela Forner, Henning Muller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Information Access Evaluation. Multilinguality, Multimodality, and Visualization, volume 8138 of Lecture Notes in Computer Science, pages 353– 367. Springer Berlin Heidelberg, 2013.
7. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):540-3.
8. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2015 May 1. pii: bbv024.
9. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P. Language Resources for French in the Biomedical Domain. Language and Resource Evaluation Conference, LREC 2014. 2014:2146-2151.
10. 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 - BioTxtM2014. 2014:24-30
11. Bodenreider, O. and McCray, A. (2003). Exploring semantic groups through visual approaches. Journal of Biomedical Informatics, 36:414–432
12. BRAT Standoff Annotation format. `http://brat.nlplab.org/standoff.html` Accessed: 2015-05-13.
13. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012.
14. Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, John-Paul Plazzer (2013) Annotating the Biomedical Literature for the Human Variome. Database: The Journal of Biological Databases and Curation, virtual issue for BioCuration 2013 meeting. 2013:bat019. doi:10.1093/database/bat019

15. Lina F. Soualmia, Chloé Cabot, Badisse Dahamna and Stéfan J. Darmoni (2015) SIBM at CLEF e-Health Evaluation Lab 2015. CLEF 2015 Online Working Notes. CEUR-WS.
16. Zubair Afzal, Saber A. Akhondi, Herman van Haagen, Erik Van Mulligen and Jan A. Kors (2015) Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. CLEF 2015 Online Working Notes. CEUR-WS.
17. Jingchi Jiang, Yi Guan and Chao Zhao (2015) WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF. CLEF 2015 Online Working Notes. CEUR-WS.
18. Maryna Chernyshevich and Vadim Stankevitch (2015) IHS-RD-BELARUS: Clinical Named Entities Identification in French Medical Texts. CLEF 2015 Online Working Notes. CEUR-WS.
19. Eva D'Hondt, François Morlane-Hondère, Léonardo Campillos, Dhouha Bouamor, Swen Ribeiro and Thomas Lavergne (2015) LIMSI @ CLEF eHealth 2015 - task 1b. CLEF 2015 Online Working Notes. CEUR-WS.
20. Jorge Vivaldi, Horacio Rodriguez and Viviana Cotik (2015) Semantic tagging of French medical entities using distant learning. CLEF 2015 Online Working Notes. CEUR-WS.
21. Devanshu Jain (2015) Supervised Named Entity Recognition for Clinical Data. CLEF 2015 Online Working Notes. CEUR-WS.