The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts

Ander Intxaurrondo[1], Martin Pérez-Pérez[2], Gael Pérez-Rodríguez[2], Jose Antonio López-Martín[3], Jesus Santamaría[1], Santiago de la Peña[1], Marta Villegas[1,4], Saber Ahmad Akhondi[5], Alfonso Valencia[4], Analia Lourenço[2], Martin Krallinger[1*]

[1]Centro Nacional de Investigaciones Oncológicas (CNIO),
[2]ESEI - Department of Computer Science, University of Vigo,
[3]Hospital 12 de Octubre - Madrid,
[4]Barcelona Supercomputing Center (BSC),
[5]Elsevier Content & Innovation

Abstract. Healthcare professionals are generating a substantial volume of clinical data in narrative form. As healthcare providers are confronted with serious time constraints, they frequently use telegraphic phrases, domain-specific abbreviations and shorthand notes. Efficient clinical text processing tools need to cope with the recognition and resolution of abbreviations, a task that has been extensively studied for English documents. Despite the outstanding number of clinical documents written worldwide in Spanish, only a marginal amount of studies has been published on this subject. In clinical texts, as opposed to the medical literature, abbreviations are generally used without their definitions or expanded forms. The aim of the first Biomedical Abbreviation Recognition and Resolution (BARR) track, posed at the IberEval 2017 evaluation campaign, was to assess and promote the development of systems for generating a sense inventory of medical abbreviations. The BARR track required the detection of mentions of abbreviations or short forms and their corresponding long forms or definitions from Spanish medical abstracts. For this track, the organizers provided the *BARR medical document collection*, the *BARR corpus* of manually annotated abstracts labelled by domain experts and the *BARR-Markyt* evaluation platform. A total of 7 teams submitted 25 runs for the two BARR subtasks: (a) the identification of mentions of abbreviations and their definitions and (b) the correct detection of short form-long form pairs. Here we describe the BARR track setting, the obtained results and the methodologies used by participating systems. The BARR task summary, corpus, resources and evaluation tool for testing systems beyond this campaign are available at: `http://temu.inab.org`.

1 Introduction

There is an increasing adoption of electronic health records (EHRs) in the European Union, promoted both by national plans as well as European initiatives like the 2 billion Euro public-private partnership Innovative Medicines Initiative [18]. Electronic

health records encompass a substantial amount of unstructured clinical texts, a key information source for clinical decision support (CDS), patient cohort stratification, and disease/adverse drug event surveillance or population health management [8, 26]. For instance, according to estimates provided by the Galician Healthcare Service, approximately 80 percent of the content of EHRs generated during the last years by this Spanish region is available as unstructured data, i.e. clinical texts. When considering the amount of unstructured clinical information generated only by the Galician health system, about 200.000 clinical notes are produced on average just on a single day. Enabling a better exploitation of the information contained in clinical notes would empower mechanisms for improving patient assistance.

The expansion of EHRs within healthcare systems promoted the development of clinical natural language processing techniques. Such systems have the aim to assist in the process of transforming clinical text written by healthcare professionals into structured clinical data representations. Clinical text mining and natural language processing (NLP) systems have been applied to a considerable number of different tasks [26, 8, 46]. Among the most prevalent clinical text processing tasks are (a) automated clinical coding [10], (b) automatic de-identification/anonymization of clinical notes[45], (c) recognition of clinical entities or concepts in running text [19], (d) negation detection [14], (e) experiencer detection (subject identification) [11], (f) temporal status (temporality) classification [42], (g) coreference resolution/anaphoric relations extraction[51, 37] or (h) detecting adverse drug events [9]. Although clinical text is a very abundant type of health data, it is also the most arduous to explore computationally. There are several inherent particularities underlying clinical texts that cause important complications to automatic text processing attempts not encountered in well-written prose or scientific literature. Clinical notes are generally far less structured and frequently do not follow normal grammar, containing ungrammatical expressions, lack of punctuation marks and accentuation as well as presence of conjoined words. Due to time constraints, clinical writings frequently contain spelling and typing errors together with author- and domain-specific idiosyncratic, often cryptic expressions.

Despite the heavy use of formal and domain specific terminologies in clinical notes, these are generally used in a rather informal and unsorted way. Moreover, in Spanish clinical texts, some studies also point out issues related to the incorrect use of Spanish medical expressions due to false friends and wrong translations from English medical terms into Spanish [4]. Benavent and Iscla describe commonly encountered cases of incorrect use of Spanish medical language [4].

Processing of clinical texts heavily relies on the performance of pre-processing modules such as tokenization, spell checking and sentence boundary recognition. Another key component for almost any clinical text processing task is the correct identification and resolution of abbreviations. Acronyms can be seen even in Roman and Greek inscriptions, and were commonly used during the Roman Empire and the middle Ages.

Abbreviations, acronyms and symbols constitute a widely used strategy to produce a more reduced and compact representation of written medical expressions [4]. Therefore, clinical narratives show a heavy use of shorthand lexical units, abbreviations and acronyms, including cases of local or even misspelled abbreviations. The correct interpretation of abbreviations is a challenge even for health care practitioners themselves

and can potentially result in medical errors due to wrong interpretations of highly ambiguous cases [43, 12]. Some estimates on English biomedical texts showed that acronyms are overloaded 33% of the time, and often correspond to highly ambiguous cases, even given contextual information [25].

Abbreviations are heavily used in Spanish EHRs, as highlighted by several studies both in Spain as well as in Latin America. For instance, Plasencia and Moliner found almost 22 abbreviations per record when examining nursing notes, discharge and emergency discharge reports [32]. They also pointed out that many of the examined abbreviations did have more than one potential meaning or interpretation and that in some cases abbreviations where wrongly used. In another study carried out by Benavent and colleagues, on average a total of 14,7 abbreviations were detected per document when dealing with emergency notes, discharge reports and clinical reports from specialized healthcare services [31]. Abbreviations are also very common in EHRs written in English, as described by [48], which detected on average 35 abbreviations in discharge summaries from the Vanderbilt Medical Center.

Abbreviations are being widely used in scientific texts, not only in case of English documents but also in Spanish medical literature. Especially publications belonging to biomedical and clinical disciplines are particularly overloaded with abbreviations and acronyms [23].

Clinical texts generally do not contain explicit mentions of contractions or abbreviations (short forms) together with their corresponding full versions or descriptive forms (long forms), also known as abbreviation-definition pairs. In order to be able to resolve abbreviations and to build sense inventories the availability of resources covering abbreviations and their definitions is critical. Although many lexical resources do exist for English biomedical abbreviations, only few manually constructed lexical resources have been generated for Spanish that might serve as resources to interpret abbreviations. These are designed mainly for human consumption and are generally not distributed in machine-readable formats. Navarro presented a manually constructed resource of abbreviations and acronyms commonly used in Spanish medical texts that serves as a valuable aid for interpreting abbreviations [28, 49]. Also lexical resources such as the Unified Medical Language System (UMLS) have been explored to extract automatically a considerable number of abbreviation definitions, i.e. abbreviation - full form pairs [22].

Another important resource for abbreviations is the medical literature. Often key elements of clinical studies can in fact also be found in medical abstracts [7]. Many scientific publications do require that authors provide a definition of abbreviations and acronyms the first time they are used in the text. According to some estimates, using English scholarly publications, 25 percent of abbreviations were explicitly defined in biomedical articles [50]. When examining different manually annotated abbreviation corpora for English biomedical abstracts, on average between 0.80 to 1.43 abbreviation definitions could be found per abstract [16].

The recognition and disambiguation of biomedical abbreviations is an intensively studied research topic in English [44]. A range of different methods have been applied to address this problem, including alignment-based approaches described by [39], machine learning techniques tested by [5], or rule-based approaches explored by [3].

In order to be able to evaluate and develop abbreviation recognition tools for English biomedical texts, several manually annotated corpora have been constructed, i.e. the MEDSTRACT, Ab3P, BOADI and Schwartz and Hearst corpora (see [17] and [40] for more details).

Unfortunately, far less research has been performed on Spanish medical abbreviation recognition, despite the volume of EHR written in that language and the existence of a considerable number of Spanish medical publications. Rubio-Lopez et al. can be counted among the few published attempts that handled acronym disambiguation in Spanish EHRs [35]. This is partially due to the lack of annotated resources and available corpora covering abbreviation annotations.

A mechanism to promote the development of biomedical/medical text mining and natural language processing systems, as well as to determine the state of the art techniques and performance to address a particular task is through shares tasks and community challenges [13]. A considerable number of challenges have been organized so far for English biomedical texts, serving as an important driving force for generating the necessary resources for the implementation of medical text processing systems [13].

We have thus organized a track specifically devoted to the automatic processing of Spanish medical literature, focusing on an important building block task, namely the automatic detection of abbreviations. The Biomedical Abbreviation Recognition and Resolution (BARR) track had the aim to promote the development and evaluation of biomedical abbreviation identification systems by providing Gold Standard training, development and test corpora manually annotated by domain experts with abbreviation-definition pairs within abstracts of biomedical documents written in Spanish.

The proposed Biomedical Abbreviation Recognition and Resolution (BARR) track has the aim to promote the development and evaluation of biomedical abbreviation identification systems by providing Gold Standard training, development and test corpora manually annotated by domain experts with abbreviation-definition pairs within abstracts of biomedical documents written in Spanish.

This paper describes the data used to support the BARR track, as well as the results obtained for this track. Section 2 describes BARR track setting and posed tasks. Section 3 provides a sort summary of the corpus and resources provided for the BARR track. In section 4 we give a brief explanation of the Markyt benchmark platform and the used evaluation measures. In section 5 we focus on the methods used by the participants while section 6 provides an overview of the obtained results. Finally, section 7 offers concluding remarks.

2 Task Description

The BARR track was one of the five tasks of the IberEval 2017 evaluation campaign, which had the aim to promote the development of language technologies for Iberian languages. The purpose of the BARR track was to explore settings that are relevant for processing both medical texts as well as clinical research narratives. The underlying assumption here was that techniques tailored to medical literature could be potentially adapted for processing clinical texts [26]. Figure 1 provides a general overview of the BARR task setting.

In essence, the BARR track evaluated systems that are able to detect mentions of abbreviation-definition pairs, i.e. short form-long form mentions that co-occurred within sentences found in abstracts of Spanish medical articles. This implied that, instead of requesting the detection and resolution of all the abbreviations found in an abstract, in this first edition of BARR, the focus was only on the discovery of abbreviation that were explicitly defined through their long forms in the same sentence. In line with some of the previously proposed resources we refer to an abbreviation as a Short Form (SF) that is a shorter term that denotes a longer word or phrase. On the other hand, the definition (the Long Form, LF) refers to the corresponding definition found in the same sentence as the SF.

The BARR task was divided into two separate subtasks, which were carried out on the same datasets [20]. The first subtask focused on the detection of mentions, in terms of their corresponding character offsets, of both short forms as well as long forms. This implied that participating teams had to detect correctly the start and end indices corresponding to all the short forms and long forms mentioned in titles and abstracts of Spanish medical articles. The second, and main subtask, focused on the detection of the actual pairs of short-form/long-from mentions in running text.

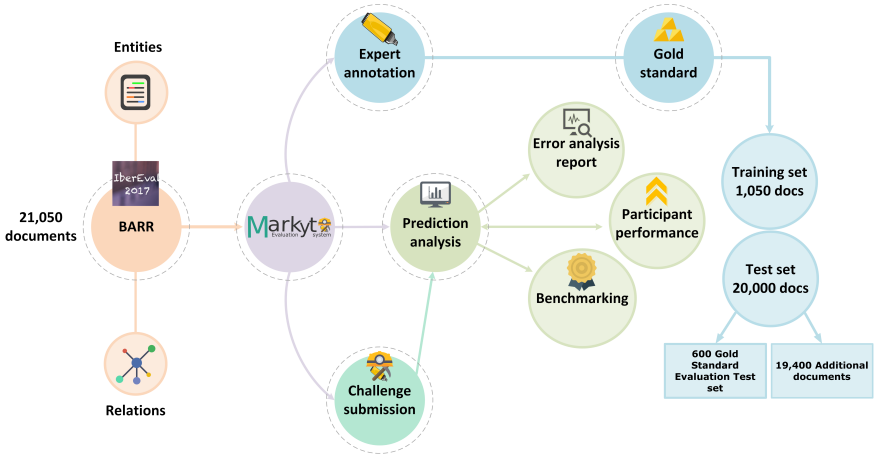

Fig. 1. BARR track overview

Participating systems were provided with a training set to construct their predictor or system during the training phase [20]. All abbreviations used for the training and test set collections were generated through an exhaustive manual annotation by domain experts, following well-defined annotation guidelines. At a later stage, a blinded test set was released for which they were asked to submit predictions that were evaluated against manual annotations. For evaluation purposes we only examined exact matches of automatically produced annotations against manual ones. For the BARR track we used a particular setting related to the test set release, similar to the evaluation scenario used for the CHEMDNER for chemical named entity recognition systems [21].

Abstracts	**Train set**	**Background 1**	**Background 2**	**Test set**	**Evaluation set**	**Elsevier**
Total number of tokens	234210	24893659	32284138	4463620	132998	7390479
Maximum number of tokens	704	1372	1372	971	642	1192
Minimum number of tokens	21	1	1	3	31	1
Mean number of tokens	223	220	208	223	221	176
Variance	9511.68	9813.74	12235.06	9799.92	9020.72	17367.18
Standard deviation	97.53	99.06	110.61	98.99	94.98	131.78
Total number of tokens in abstracts (complete corpus): 36981968						
Titles	**Train set**	**Background 1**	**Background 2**	**Test set**	**Evaluation set**	**Elsevier**
Total number of tokens	14015	1499944	2039335	306166	8996	539391
Maximum number of tokens	43	104	154	56	41	154
Minimum number of tokens	2	1	1	3	4	1
Mean number of tokens	13	13	13	15	14	12
Variance	33.90	35.77	36.85	35.54	34.85	40.60
Standard deviation	5.82	5.98	6.07	5.96	5.90	6.37
Total number of tokens in titles (complete corpus): 2359516						

Table 1. BARR corpus statistics.

The documents released during the test phase included, in addition to the Gold Standard evaluation test set used to assess the performance of participating systems, an additional larger collection of documents to explore robustness and scalability of the systems and to make sure that any manual revision or correction of results prior to submission would be unfeasible. Each participating BARR teams was allowed to submit for each of the sub-tasks a total of up to five predictions (runs). In order to coordinate the participation and evaluation of teams of the BARR track, the BARR task organizers offered, in addition to a general task website, an evaluation and benchmark platform adapted to this task, the BARR-Markyt platform [30]. This platform enables user registration and notification, as well as management, evaluation, visualization and benchmarking of systems and their annotations. Systems had to upload their predictions to the BARR-Markyt platform in a predefined format to be considered for evaluation.

3 Data Sets and Resources

A considerable barrier for the development of Spanish medical text processing systems is the lack of a unified repository or publication aggregator of all medical and biomedical literature published in Spanish. Note that the centralized citation repository PubMed does provide article abstracts only in English for a large number of articles written originally in Spanish.

The underlying scenario for Spanish medical literature is rather fragmented, with multiple (partially overlapping) resources offering abstracts and/or publications of medical literature written in Spanish [2]. In order to provide a large document collection of Spanish medical abstracts for the BARR track, going beyond the actual training and test documents, we constructed the BARR document collection by integrating abstracts and titles from multiple sources including a specially constructed set of records provided by the publisher Elsevier specifically for this track. The BARR document or background collection contained a total of 155,538 records, out of which Elsevier provided 41,760 documents, while the rest corresponded to publications combined from multiple different sources. To distinguish these two collections they will be referred to as BARR

background 1 (background set without the Elsevier collection) and background 2 (entire background set including the Elsevier collection).

A subgroup from the BARR document collection was used to construct a manually labelled training and test set. The training and the test set were random subsamples from the same document set to avoid selection bias during the evaluation step. The entire BARR training set contained 1,050 abstracts, while the BARR Gold Standard test set had a total of 600 abstracts.

Table1 provides a statistical overview of the BARR datasets, covering basic corpus statistics of the abstracts and titles of the training set, background sets, and the test set; we also included information about Elsevier BARR documents alone.

The BARR document corpus was released in form of plain-text, UTF8-encoded abstracts in a simple tab-separated format with columns corresponding to the document identifier, an ISO 639-1 language two-letter code corresponding to the language of the abstract body, title of the record and abstract of the record.

The manual labelling of abbreviation mentions of the BARR corpus was done using a customized version of AnnotateIt. During a second follow up annotation step, a customized version of the Markyt annotation system was used to manually revised mention annotations and to annotate the relations between short forms and long forms (as well as short forms and nested mentions) co-mentioned in sentences [30].

To provide a richer annotated corpus, not limited only to SF-LF pairs, the BARR corpus provided annotations for a total of eight different abbreviation-related mention types, which are summarized in figure 2 (A) together with an example abstract (B). All annotations were annotated by biomedical experts and supervised by a practising oncologist with an additional degree in bioinformatics and basic knowledge in text mining.

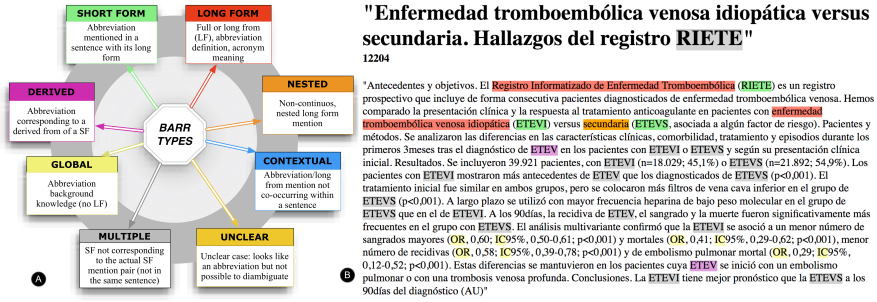

Fig. 2. BARR annotation types

Table 2 shows the frequency of entity types and relation types in the training set and the background set organized by titles and abstracts.

The average frequency of abbreviations found per abstracts depended heavily on the considered mention type. For instance short forms occurred in the training set 1.41 times per abstract, while multiple mentions occurred 2.84 times and global abbreviations (those missing an explicit long form in the abstract) occurred on average 3.37 times.

Entity type	Train		Test		Relation type	Train		Test	
	Title	Abstract	Title	Abstract		Title	Abstract	Title	Abstract
LONG	28	892	10	505	SF - LF	28	879	9	491
SHORT	27	915	10	517	NESTED - SF	0	47	0	16
MULTIPLE	25	1878	13	1144	LF - NESTED	0	48	0	15
GLOBAL	133	2124	71	1279					
CONTEXTUAL	35	199	14	52					
UNCLEAR	3	68	0	7					
DERIVED	1	65	0	30					
NESTED	0	45	0	21					

Table 2. Manually annotated entities and relations in the BARR corpus, organized by titles and abstracts.

A more exhaustive examination of the BARR document collection showed that a small set of abstracts did correspond to records that were in fact not written in Spanish but in another language, mostly English. Therefore all the records in the training and test set were manually classified into its corresponding language, being only the abstracts written in Spanish considered for evaluation purposes, while for the background sets, an automatic language detection algorithm was used to assign language codes.

Finally, the task organizers provided a list of additional resources including tutorials, results of available abbreviation extraction systems adapted to Spanish and a collection of software, datasets, and lexical resources relevant to the BARR track.

4 Evaluation Measures

The Markyt web-based benchmarking platform supported the evaluation of the predictions of the participating teams [30]. The primary evaluation metric used for the BARR track consisted in micro-average F-measure.

Markyt provided micro/macro-average standard performance statistics, such as recall, precision and F-score, and enabled the examination of annotation mismatches (see [41] and [33]).

Correspondingly, recall (Eq. 1) is the percentage of correctly labelled positive results over all positive cases, being a measure of the ability of a system to identify positive cases.

$$recall = \frac{TP}{TP + FN} \tag{1}$$

Precision (Eq. 2) represents the percentage of correctly labelled positive results over all positive labelled results, i.e. it is a measure of the reproducibility of a classifier of the positive results.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

Lastly, F-score (or balanced F-measure), the primary evaluation metric, stands for the harmonic mean between precision and recall (Eq. 3).

$$F - Score = \frac{2 * (precision * recall)}{precision + recall} \tag{3}$$

Micro-average statistics were calculated globally by counting the total true positives, false negatives and false positives. Conversely, macro-average statistics were calculated by counting the true positives, false negatives and false positives on a per-document basis and then, averaged across documents.

During the test phase, teams were requested to generate predictions for a blinded collection of documents, and they had to upload their submission to the BARR-Markyt system within a short period of time. Teams could submit up to five prediction files (runs). Additionally, three main result types were examined: false negative (FN) results corresponding to incorrect negative predictions (i.e. cases that were part of the gold standard, but missed by the automatic system), false positive (FP) results being cases of incorrect positive predictions (i.e. wrong results predicted by the automatic system that had no corresponding annotation in the gold standard) and true positive (TP) results consisting of correct positive predictions (i.e. correct predictions matching exactly with the gold standard annotations). The micro-averaged recall, precision and F-score statistics were used for final prediction scoring, and F-score was selected as main evaluation metric. Figure 3 illustrates schematically the submission process using Markyt.

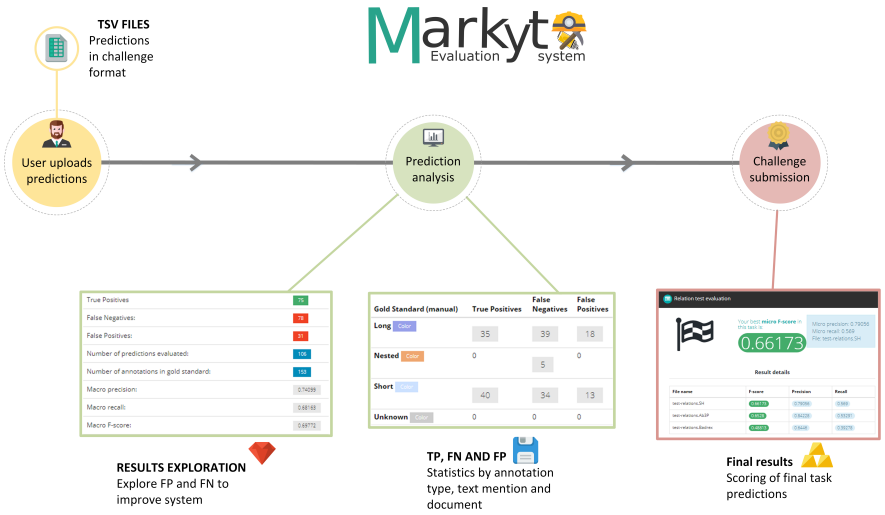

Fig. 3. BARR participant submission overview

5 Overview of the Submitted Approaches

A total of 17 teams registered for the BARR track through the BARR-Markyt team registration page, whereof 7 teams returned correctly submission before the team results submission due. The evaluated BARR teams submitted a total of 25 runs, 13 for the subtask of recognizing abbreviation mentions and 12 for the detection of mentions of short form-long form pairs. Below we summarize briefly the most relevant aspects of

the used methodology by participating teams, while additional details can be found in the BARR team technical note papers [15, 27, 34, 36, 1].

The *IBI-UPF team* used an approach structured into two sequential steps, namely an entity spotting phase and a relation extraction phase [34]. For the frist step they trained three token-based Random Forest classifiers that modelled both the token itself as well as a context window of 2 characters: (1) an abbreviation token classifier, (2) a long form token classifier and (3) an abbreviation type classifier, i.e. whether the abbreviation is a short form or another type of abbreviation. During the second phase, they applied a set of heuristics and a scoring function to identify SF-LF pair relations.

The system used by the *UC-III team* was based on a rule-based approach [36]. They implemented a technique consisting of an adaptation of an algorithm originally proposed by Schwartz and Hearst [38] to handle English biomedical abbreviations and tailored it to Spanish texts together with additional modifications. Their strategy assumed that long forms are mentioned in the same sentence and before their corresponding short forms. Abbreviations were extracted through pattern rules, imposing maximum length cut-offs to identify the short forms as well as long forms. Short forms had to start with an alphanumeric character and had to contain at least one letter. Moreover, the first character of the short form had to match the character of the initial position of the first word of the detected long form. The system of UC-III team was particularly fast both for the abbreviation recognition as well as for the SF-LF relation detection.

The *IXA-UPV/EHU team* addressed this task as a classical NER scenario by applying two standard, domain agnostic, machine-learning based NER taggers for the recognition of abbreviation mentions [1]. They managed to submit results for one of their systems, namely a perceptron tagger based on sparse, shallow features. Their system relied on the IXA pipes tokenizer for pre-processing and tokenization, together with the ixa-pipe-nerc for short form and long form mention recognition. This team used the BARR background set to induce clusters and word embeddings for training both of their systems. For the ixa-pipe-nerc tagger they used a customized version of the Apache OpenNLP implementation of the perceptron algorithm.

The *CNIO team* provided a comparative benchmarking of three publicly available state-of-the-art biomedical abbreviation detection and recognition systems. These systems, based on heuristics, were originally developed for English biomedical texts and thus had to be adapted to handle Spanish documents. Tailoring to Spanish covered essentially processing of accentuated characters. The three approaches that were benchmarked by this team included the Ab3P, ADRS and BADREX taggers. These tools exploit heuristics related to the presence of parentheses and the detection of abbreviation candidates within them, together with the exploration of the nearby context for detecting SF-LF pairs. These taggers analyse characters inside parentheses, and detect words that correspond to long form candidates by matching these characters in certain positions.

The *EHU team* adapted a system that relied on the FreelingMed analyser to find abbreviations [27]. This system incorporated a dictionary contained in FreelingMed that comprised abbreviations and their expanded forms appearing in SNOMED CT. Furthermore, this team also used heuristics to recognize word-forms based on particu-

lar patterns that were indicative for abbreviations. Finally, through Freeling they could mark abbreviations referring to units of weight, length or time.

The *UPC team* used three different tactics for detecting acronym-expansion pairs [27]. Their similarity-based strategy relied on 13 hand crafted extraction rules that were applied in decreasing order of confidence. These rules took into account aspects such as word shape features observed in training set annotations. Their gazetteer-based strategy relied on a large medical terminological resource integrating entries from multiple sources. This terminology contained 14,360 short forms-long form pairs that were detected in the terminology using the similarity-based method. Finally, their distance-based strategy was based on a list of patterns covering acronym-expansion pairs co-occurring closely and frequently in sentences of the training data.

The *UNED team* participated through a two-step approach that identified first the abbreviation candidates and then tried to match potential definitions [27]. For the identification of abbreviations they essentially considered upper case terms, terms with combinations of upper case letters and other characters as well as the presence of parenthesis. Also this team employed an adaptation of the Schwartz and Hearst algorithm for detecting SF-LF pairs in running text [38]. Their extension of this algorithm incorporated cases where the characters of the short form did not appear in the identical order than the corresponding long form matching characters. Moreover, the number of words of long forms was not allowed to exceed the double of the characters of its corresponding short form. They also constructed a set of patterns to handle frequent special cases. When the pattern based-strategy did not detect a valid definition they used a dictionary of 7,916 entries to match possible long forms.

The last three teams (EHU, UPC and UNED) are described in a joined working notes paper exploring output combinations of these strategies [27]. According to their conclusion, combinations did not improve substantially the results when compared to the single best system from the UNED team.

6 Results and Discussion

We received a total of 25 runs from the seven teams that submitted results for the subtasks of the BARR track. Dividing the BARR track into two sub-task allowed us to carry out a more granular evaluation analysis and enabled participants to improve their systems precisely for each of the underlying components.

We evaluated a total of 13 runs for the first sub-task, the detection of mention offsets of abbreviation short forms and abbreviation definitions, i.e. long forms and nested long forms. Table 3 illustrates the obtained performance for each of the evaluated submissions for this sub-task. The top scoring prediction was provided by the UNED team, obtaining an F-measure of 74.91%, followed by the UC-III team with an F-measure of 72.05%. The UC-III team, which relied on a customized version of the Schwartz and Hearst algorithm, obtained the highest recall, corresponding to 73.47%. The second highest recall was obtained by run 4 of the IBI-UPF team (69.85%). All submissions showed a consistently higher precision when compared to the recall results. The top scoring precision was reached by the AB3P benchmarking run of the CNIO team

Abbreviation team runs	Precision	Recall	F1-score
UNED	86.84	65.86	**74.91**
UC-III	70.69	**73.47**	72.05
CNIO-ADRS	84.74	62.29	71.7
IBI-UPF-v4	72.2	69.85	71
IBI-UPF-v3	71.88	69.53	70.69
CNIO-AB3P	**87.95**	56.72	68.96
IBI-UPF-v1	67.21	68.69	67.95
IXA-UPV/EHU	71.78	59.87	65.29
CNIO-BADREX	81.60	47.06	59.69
UNED-or	76.69	37.29	50.17
UPC	65	21	32
UNED-and	99	10	18
EHU	25	10	15

Table 3. Results of the BARR abbreviation mention recognition runs.

(87.95%) followed by the UNED submission (86.84%). These results indicate that rule based approaches did obtain a competitive performance for this subtask.

In summary, the achieved results indicate that participating systems were able to obtain a reasonable performance for this subtask. Nonetheless, these results also suggest that there is room for additional performance improvement. These results also imply that this task was considerably more difficult when compared to analogous approaches tested on English biomedical abstracts. In order to determine potential aspects affecting the abbreviation mention recognition systems performance we carried out a detailed analysis of annotations that were often wrongly predicted across multiple runs. When differentiating the results of different mention types, it became clear that short forms were considerable easier to detect when compared to long forms or nested long form mentions. Many of the errors in long form recognition referred to incorrect mention boundary detection, particularly in cases of very long definitions. The presence of hyphens, numbers, certain non-alphanumeric characters or accentuated characters was also observed frequently within missed mentions, of both short forms and long forms. Long forms with conjoined words were another difficult case for participating teams.

A frequently missed type of short form corresponded to single letter abbreviations (e.g. T - telaprevir) or abbreviations that contained punctuation marks (e.g. P.M.M.). Another particularly difficult type of abbreviation, also affecting the SF-LF relation extraction subtask, corresponded to the special case of non-Spanish short forms, whereas the corresponding co-mentioned long form description was written in Spanish. This special scenario is particularly widespread in the biomedical and medical literature due to the influence of English academic and technical terminology. Most of the non-Spanish abbreviations corresponded to English terms, while a minor fraction corresponded to abbreviations in another language such as Latin. This special "bi-lingual short form-long form pairs" were kept due to their practical relevance. For instance abbreviations such as PSA (prostate specific antigen) is more often used in the Spanish medical texts as abbreviation for *antígeno prostático específico* instead of its Spanish

Relation team runs	Precision	Recall	F1-score
UNED	81.53	60.93	**67.74**
UC-III	72.21	**61.78**	66.59
CNIO-ADRS	79.05	56.9	66.17
CNIO-AB3P	84.23	53.29	65.28
IBI-UPF-v4	71.82	50.32	59.17
IBI-UPF-v3	71.21	49.89	58.67
IBI-UPF-v1	63.3	47.98	54.59
CNIO-BADREX	64.46	39.28	48.81
UNED-or	57.53	26.75	36.52
UPC	**97.67**	8.91	16.34
UNED-and	31	10	15

Table 4. BARR abbreviation relation evaluation.

counterpart (APE). The detection of this kind of SF-LF pairs would require the use of either bilingual abbreviation definition lexical resources or even machine translation techniques.

The second, and main subtask of the BARR track, focused on the detection of pairs of short-form/long-from mentions in running text. This task was obviously directly dependent on the results of the previous sub-task. Table 4 provides the evaluation results of the performance of all received submissions for this sub-task.

The run of the UNED team obtained the highest F-measure (67.74%) followed by the UC-III submission (66.59%). The UC-III run reached the top scoring recall (61.78%) whereas the UNED run obtained the second highest recall score (60.93%). When examining precision, the UPC run obtained the highest score (97.67%), but with a very low recall (recall 8.91%). The CNIO-AB3P obtained the second highest precision score (84.23%). An examination of prediction errors for the SF-LF pair task showed that many mismatches were caused by incorrect boundary recognition of the long form portion. We observed among the types of frequently missed SF-LF pairs, relations where the short form was stated in the text before the long form (and not afterwards). Other difficult cases corresponded to pairs where several characters of the short form appeared in a single word of its long form (e.g. CTP - catéter telescopado, or CMV cytomegalovirus). Long forms with permuted word order not following its short form character order were problematic for many teams, while in some cases also pairs that contained SFs or LFs with accentuated characters were not detected correctly. Curiously SF-LF pairs corresponding to substances and chemical compounds seemed to imply some difficulty. Finally, cases where parentheses serving as abbreviation markers were missing caused also in a drop in performance.

7 Conclusions

The BARR track was able to promote the development of resources, corpora and processing tools for a key task of medical text mining, the recognition of abbreviations. The BARR track was promoted by the Plan for the Advancement of Language Technology,

a Spanish national plan to encourage the development of natural language technologies for Spanish and Iberian languages [47]. The obtained results highlight that participating teams were able to implement systems that can be valuable for the development of lexical resources for disambiguating abbreviations. Improvement of future systems might require access, not only to additional corpora, but also to better basic medical language processing and pre-processing components. Follow up tasks should address the disambiguation of all abbreviations beyond SF-LF pairs, including also other document types such as clinical notes as well as identifying the knowledge domain to which an abbreviation belongs [24]. Several teams continued the improvement of their systems after the test phase. In order to encourage further improvements and the implementation of new systems beyond the BARR test phase we plan the construction of an additional blinded Gold Standard dataset that can be used to assess future tools through the BARR-Markyt system. We will also publish the results of the BARR inter-annotator agreement study on the track web. We expect that the BARR corpus should be distributed in alternative formats such as the popular BioC format [6]. Future tasks should also align with interoperability aspects to facilitate integration of the tools of participating teams into platforms such as the OpenMinted infrastructure[1] in addition to technical assessment as done by benchmarking initiatives like BeCalm[2] [29].

8 Acknowledgments

We acknowledge the Encomienda MINETAD-CNIO/OTG Sanidad Plan TL and OpenMinted (654021) H2020 project for funding.

References

1. Agerri, R., Rigau, G.: Applying existing named entity taggers at barr ibereval 2017 task. SEPLN (2017)
2. Aleixandre-Benavent, R., Bolaños-Pizarro, M., de Dios, J.G., Navarro-Molina, C.: Fuentes de información bibliográfica (ii). bases de datos bibliográficas españolas en ciencias de la salud de interés en pediatría: Ime, ibecs y medes/sources of bibliographic information (ii). spanish bibliographic database in health sciences of interest in pediatrics. databases of ime, ibecs and medes. Acta Pediatrica Espanola 69(4), 177 (2011)
3. Ao, H., Takagi, T.: ALICE: an algorithm to extract abbreviations from MEDLINE. Journal of the American Medical Informatics Association : JAMIA (2005)
4. Benavent, R.A., Iscla, A.A.: Problemas del lenguaje médico actual.(ii) abreviaciones y epónimos. Papeles Méd 10(4), 170–6 (2001)
5. Chang, J.T., Schtze, H., Altman, R.B.: Creating an online dictionary of abbreviations from medline. Journal of the American Medical Informatics Association (2002)
6. Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., et al.: Bioc: a minimalist approach to interoperability for biomedical text processing. Database 2013, bat064 (2013)

[1] http://openminted.eu/
[2] http://www.becalm.eu/

7. Dawes, M., Pluye, P., Shea, L., Grad, R., Greenberg, A., Nie, J.Y.: The identification of clinically important elements within medical journal abstracts: Patient_population_problem, exposure_intervention, comparison, outcome, duration and results (pecodr). Journal of Innovation in Health Informatics 15(1), 9–16 (2007)
8. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? Journal of biomedical informatics 42(5), 760–772 (2009)
9. Friedman, C.: Discovering novel adverse drug events using natural language processing and mining of the electronic health record. Artificial Intelligence in Medicine pp. 1–5 (2009)
10. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association 11(5), 392–402 (2004)
11. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics 42(5), 839–851 (2009)
12. on Accreditation of Healthcare Organizations, J.C., et al.: Medication errors related to potentially dangerous abbreviations. Sentinel Event Alert 23, 1–4 (2001)
13. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics 17(1), 132–144 (2015)
14. Huang, Y., Lowe, H.J.: A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of the American Medical Informatics Association 14(3), 304–311 (2007)
15. Intxaurrondo, A., Krallinger, M.: Cnio at barr ibereval 2017: exploring three biomedical abbreviation identifiers for spanish biomedical publications. SEPLN (2017)
16. Islamaj Doan, R., Comeau, D.C., Yeganova, L., Wilbur, W.J.: Finding abbreviations in biomedical literature: three bioc-compatible modules and four bioc-formatted corpora. Database : the journal of biological databases and curation 2014 (2014)
17. Islamaj Doan, R., Comeau, D.C., Yeganova, L., Wilbur, W.J.: Finding abbreviations in biomedical literature: three bioc-compatible modules and four bioc-formatted corpora. Database (2014)
18. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 13(6), 395–405 (2012)
19. Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., Xu, H.: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association 18(5), 601–606 (2011)
20. Krallinger, M., Intxaurrondo, A., Lopez-Martin, J., de la Peña, S., Pérez-Pérez, M., Pérez-Rodríguez, G., Santamaría, J., Villegas, M., Akhondi, S., Lourenço, A., Valencia, A.: Resources for the extraction of abbreviations and terms in spanish from medical abstracts: the barr corpus, lexical resources and document collection. SEPLN (2017)
21. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., Sayle, R.A., Batista-Navarro, R.T., Rak, R., Huber, T., Rocktschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K.H., Ramanan, S.V., Nathan, S., itnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S.A., Kors, J.A., Xu, S., An, X., Sikdar, U.K., Ekbal, A., Yoshioka, M., Dieb, T.M., Choi, M., Verspoor, K., Khabsa, M., Giles, C.L., Liu, H., Ravikumar, K.E., Lamurias, A., Couto, F.M., Dai, H.J., Tsai, R.T.H., Ata, C., Can, T., Usi, A., Alves, R., Segura-Bedmar, I., Martnez, P., Oyarzabal, J., Valencia, A.: The chemdner corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics 7, S2 (2015)
22. Liu, H., Lussier, Y.A., Friedman, C.: A study of abbreviations in the umls. Proceedings. AMIA Symposium pp. 393–397 (2001)

23. Liu, H., Aronson, A.R., Friedman, C.: A study of abbreviations in medline abstracts. Proceedings. AMIA Symposium pp. 464–468 (2002)
24. Liu, H., Johnson, S.B., Friedman, C.: Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the umls. Journal of the American Medical Informatics Association : JAMIA 9, 621–636 (2002)
25. Liu, H., Lussier, Y.A., Friedman, C.: Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. Journal of biomedical informatics 34(4), 249–261 (2001)
26. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., et al.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 35(128), 44 (2008)
27. Montalvo, S., Oronoz, M., Rodríguez, H., Martínez, R.: Biomedical abbreviation recognition and resolution by prosa-med. SEPLN (2017)
28. Navarro, F.A.: Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos médicos en español. Panace 9(27) (2008)
29. Pérez-Pérez, M., Pérez-Rodrıguez, G., Blanco-Mıguez, A., Fdez-Riverola, F., Valencia, A., Krallinger, M., Lourenco, A.: Benchmarking biomedical text mining web servers at biocreative v. 5: the technical interoperability and performance of annotation servers-tips track. Proceedings of the BioCreative 5, 12–21 (2017)
30. Pérez-Pérez, M., Pérez-Rodríguez, G., Rabal, O., Vazquez, M., Oyarzabal, J., Fdez-Riverola, F., Valencia, A., Krallinger, M., Lourenço, A.: The markyt visualisation, prediction and benchmark platform for chemical and gene entity recognition at biocreative/chemdner challenge. Database 2016 (2016)
31. Piñero, L., Aleixandre-Benavent, R., Ibáñez, A.B.: Uso y abuso de abreviaturas y siglas entre atención primaria, especializada y hospitalaria. Papeles médicos 15(2), 29 (2006)
32. Plasencia Martínez, S., Moliner Lahoz, J.: Uso y abuso de las abreviaturas en los informes del hospital clínico universitario lozano blesa. Rev Paraninfo Digital 16 (2012)
33. Rijsbergen, C.J.V.: Information Retrieval. Butterworth-Heinemann (1979)
34. Ronzano, F., Furlong, L.: Ibi-upf at barr-2017: learning to identify abbreviations in biomedical literature. system description. SEPLN (2017)
35. Rubio-López, I., Costumero, R., Ambit, H., Gonzalo-Martín, C., Menasalvas, E., Rodríguez, G.A.: Acronym disambiguation in spanish electronic health narratives using machine learning techniques. Studies in health technology and informatics 235, 251 (2017)
36. Sánchez, C., Martínez, P.: A proposed system to identify and extract abbreviation definitions in spanish biomedical texts for the biomedical abbreviation recognition and resolution (barr) 2017. SEPLN (2017)
37. Savova, G.K., Chapman, W.W., Zheng, J., Crowley, R.S.: Anaphoric relations in the clinical narrative: corpus creation. Journal of the American Medical Informatics Association 18(4), 459–465 (2011)
38. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing pp. 451–462 (2003)
39. Schwartz, A., Hearst, M.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: In Proceedings of Pacic Symposium on Biocomputing (2003)
40. Sohn, S., Comeau, D.C., Kim, W., Wilbur, W.J.: Abbreviation definition identification based on automatic precision estimates. BMC bioinformatics (2008)
41. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. (2009)
42. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association 20(5), 806–813 (2013)

43. Tena Marsà, X.: El culto a las abreviaciones: idolatría o virtud. Reumatología Clínica 8(2), 54–55 (2012)
44. Torii, M., Liu, H., Hu, Z., Wu, C.: A comparison study of biomedical short form definition detection algorithms. In: Proceedings of the 1st International Workshop on Text Mining in Bioinformatics (2006)
45. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association 14(5), 550–563 (2007)
46. Velupillai, S., Mowery, D., South, B.R., Kvist, M., Dalianis, H.: Recent advances in clinical natural language processing in support of semantic analysis. Yearbook of medical informatics 10(1), 183 (2015)
47. Villegas, M., de la Peña, S., Intxaurrondo, A., Santamaría, J., Krallinger, M.: Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje. SEPLN (2017)
48. Wu, Y., Denny, J.C., Rosenbloom, S.T., Miller, R.A., Giuse, D.A., Xu, H.: A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In: AMIA annual symposium proceedings. vol. 2012, p. 997. American Medical Informatics Association (2012)
49. Yetano Laguna, J., Alberola Cuñat, V.: Diccionario de siglas médicas. SEDOM, Sociedad Española de Documentación Médica (2012)
50. Yu, H., Hripcsak, G., Friedman, C.: Mapping abbreviations to full forms in biomedical articles. Journal of the American Medical Informatics Association : JAMIA 9, 262–272 (2002)
51. Zheng, J., Chapman, W.W., Crowley, R.S., Savova, G.K.: Coreference resolution: A review of general methodologies and applications in the clinical domain. Journal of biomedical informatics 44(6), 1113–1122 (2011)