Vicomtech at BARR2: Detecting Biomedical Abbreviations with ML methods and dictionary-based heuristics

Montse Cuadros[1], Naiara Pérez[1], Iker Montoya[2], and Aitor García Pablos[1]

[1] Vicomtech, Paseo Mikeletegi 57, Donostia-San Sebastian, Spain
{mcuadros,nperez,agarciap}@vicomtech.org
[2] I+D Lanik S.A, Donostia-San Sebastian, Spain
{iker92montu}@gmail.com

Abstract. This paper presents the system developed by Vicomtech to participate in the Second Biomedical Abbreviation Recognition and Resolution (BARR2) track. For this purpose, we have used simple machine learning approaches on annotated electronic health records and the datasets provided in the track. The machine learning approaches have been tested individually and in combination with heuristics based on a dictionary of biomedical abbreviations adapted for the task.

Keywords: biomedical nlp · abbreviations · machine learning · dictionary-based approaches

1 Introduction

This paper describes Vicomtech's participation in the Second Biomedical Abbreviation Recognition and Resolution (BARR2) track of the third IberEval workshop (IberEval 2018), both in sub-tracks 1 and 2. Sub-track 1 consists in detecting only explicit occurrences of abbreviation-definition pairs. For sub-track 2, resolution of short forms must be provided regardless whether its definition is mentioned within the actual document. Both sub-tracks focus on clinical free text in Spanish.

This paper is organized as follows: Section 2 presents the two tasks in more detail; Section 3 presents our approaches to the two problems; Section 4 shows our results; finally, Section 5 contains our concluding remarks.

2 Biomedical Abbreviation Recognition and Resolution 2nd Edition (BARR2)

The Second Biomedical Abbreviation Recognition and Resolution track[1] is organized in two sub-tasks, sub-track1 and sub-track2. Both tasks require recognizing abbreviations and acronyms in Spanish clinical texts, and providing the correct definition for each recognized element. The difference between the

tasks is the number of abbreviations that each subtasks asks for and where the definitions should originate from.

Sub-track 1 requires detecting all the abbreviations for which the definitions are given explicitly in the document. Both the short form (i.e., the abbreviation or acronum) and the long form (i.e., the definition or description) must be reported. For example, for the following piece of text:

```
"... se aplicó radiofrecuencia (RF) sobre la vía accesoria auriculo-ventricular
(AV) de conduccin bidireccional. Se interrumpe la taquicardia y la preexcitación,
finalizando el procedimiento. Quedó con bloqueo de rama derecha (BRD)
..."
```

the answer should note the 3 short forms "RF", "AV", and "BRD", along with their explicit long forms "radiofrecuencia", "auriculo-ventricular", and "bloqueo de la rama derecha", respectively:

```
S1888-75462014000200009-1    SHORT_FORM    1524    1526    RF    SHORT-LONG
   LONG_FORM    1507    1522    radiofrecuencia
S1888-75462014000200009-1    SHORT_FORM    1573    1575    AV    SHORT-LONG
   LONG_FORM    1551    1571    auriculo-ventricular
S1888-75462014000200009-1    SHORT_FORM    1720    1723    BRD   SHORT-LONG
   LONG_FORM    1695    1718    bloqueo de rama derecha
```

Sub-track 2 requires detecting all the abbreviations within the document, and providing a resolution regardless their appearing explicitly in the text. The following text excerpt contains such 2 short forms, "RMN" and "MTT":

```
Se solicitó una RMN de pie izquierdo, que reveló una fractura de estrés
en el 2o MTT con callo perióstico...
```

The system developed for this sub-track should be able to find these two elements and give their long forms, "resonancia magnéitca nuclear" and "metatarso", respectively:

```
S1889-836X2015000200005-2    878    881    RMN    resonancia magnética
nuclear    resonancia magnético nuclear
S1889-836X2015000200005-2    943    946    MTT    metatarso    metatarso
```

The organization[2] has provided a sample set, a training set and a development set of the sizes shown in Table 1. The test set provided for evaluating the approaches was about 10 times bigger than the other sets, containing 2879 clinical tests, even though the submitted runs where eventually evaluated against a set of the same size as the training set.

3 Methodology

This work is a continuation of [5], where several experiments were performed for detecting and disambiguating abbreviations in electronic health records (EHR).

	Sample set	Training set	Development set	Testing set
Clinical tests	15	318	146	220
Sub-track 1	10	287	178	239
Sub-track 2	89	4,261	1,878	3,414

Table 1. Number of documents in the different sets

In this work, a small corpus of 149 EHRs was compiled manually annotated with 2,389 abbreviations and acronyms. These EHRs were provided by a local hospital and belong to different clinical specialties. Of the short forms annotated, 2 clinicians manually disambiguated two sets, one containing the 15th most ambiguous forms and the other the 30th most ambiguous forms. Finally, a dictionary of short- and long-form pairs was crafted based on [3] and the annotated corpora. The present work relies on the EHR corpora and the hand-crafted dictionary, in addition to the datasets provided by the organization of the track.

The following sections describe the approaches taken to the problems of abbreviation recognition (both in BARR2 sub-tracks 1 and 2), and of abbreviation resolution in sub-track 1 (i.e., finding the explicit long form) and sub-track 2. For the purpose of the BARR2 track, most of the effort has been put to the problem of recognition.

3.1 Abbreviation recognition

For each sub-track, we have trained several classifiers and envisaged two extra methods based on regular expressions and the hand-crafted dictionary in order to improve the recall of the machine learning approaches.

Machine Learning approach Several machine learning classifiers have been trained with Weka [4] (default settings), using the EHR dataset described above and both the BARR2 Training sets (BARR2 TS) for sub-track 1 and sub-track 2. The same very cheap features as in [5] have been used for learning the models:

- **Uppercase**: whether the token is all uppercase
- **Digit**: whether the token contains digits
- **Strange ending**: whether the token has a strange ending, where a strange ending is one that doesn't fit to the normal ones in tokens which are not abbreviations
- **Length**: token length
- **Uppercase count**: amount of uppercase characters in the token
- **Lowercase count**: amount of lowercase characters in the token
- **Vowel ratio**: amount of vowels in the token divided by its length
- **Punctuation ratio**: amount of punctuation characters in the token divided by its length

Table 2 shows the performance of the trained classifiers in the BARR2 Development set in terms of F1-measure. The first column refers to the models trained on the EHRs only; the second column refers to the models trained on the BARR2 TS only; finally, the last column shows the results on learning the classifiers on both datasets. Overall, combining both datasets yields worse results than using the BARR2 TS alone, except with the Random Fields (RF) algorithm, which is the best classifier obtained, followed by J48. Not surprisingly, the models trained on EHRs only perform the worst.

	EHRs	BARR2 TS	combined
J48	81.71	89.63	88.52
KNN1	78.97	89.69	88.42
Naive Bayes	56.44	60.05	58.49
REPtree	82.14	89.12	87.82
SMO	0.00	62.65	54.51
RF	79.45	91.14	**91.34**

Table 2. F1-measure of abbreviation recognition on the Development sets of sub-tracks 1 and 2, trained on 3 different corpora

Taking these results into account, the classifiers selected for the BARR2 competition have been J48 trained with BARR2 TS only and RF trained with the combined datasets.

Pattern-based approach (Pat) This approach consists of a set of regular expressions aiming to retrieve the abbreviations and acronyms that the ML approach does not cover. Basically, it retrieves all the strings of upper- and lowercase characters that have an uppercase character and are inside brackets. That is, this approach makes sense mainly in sub-track 1. Additionally, some tests have been carried out to try to retrieve short forms with digits too, but the results have worsened.

Dictionary-based approach (Regex) This approach is based on the dictionary introduced above and a set of rules hand-crafted after study and observation of the abbreviations in several sets of EHR and the literature. For this work, the dictionary developed in [5] has been refined taking in account the BARR2 Training and Development set examples. The final version of the dictionary contains 3447 unique pairs of biomedical short- and long-form pairs.

3.2 Abbreviation resolution for sub-track 1

Regarding sub-track 1, the system uses one or the combination of the Machine Learning approach, Pattern-based approach and Dictionary-based approach to

detect abbreviations candidates. Once the candidates are found and after checking they are surrounded by brackets, an 8th n-gram window before the abbreviation is considered as the possible definition. This possible definition is firstly checked against our dictionary, and if exists, we select it. Otherwise, a set of heuristics are considered in order to determine if the text before is the definition. The heuristics are based on: 1) the capital letters of the definition and the letters of the abbreviation in the same order or backwards, 2) the size of the definition related to the size of the abbreviation, 3) a priority of sizes definitions (3-ngrams > 2-ngram >4-ngram > 5-ngram ...). The different heuristics exclude the following ones when one is triggered. Finally if a definition is found, both abbreviation and definitions are selected and their offsets in the original clinical text are calculated.

3.3 Abbreviation resolution for sub-track 2

Regarding sub-track 2, the system uses one or the combination of the Machine Learning approach and Dictionary-based approach to detect the abbreviations candidates. For each possible candidate a definition is selected from our dictionary. Finally the offsets where the abbreviation is found in the clinical text are provided.

4 Experiments and Evaluation

Vicomtech has submitted a total of 4 systems to sub-track 1 and 4 systems to sub-track 2. The systems rely on either one of the approaches described above or their combinations. We have tested them with the Sample set firstly, but then refined them by using the BARR2 Training and Development sets. Pat and Regex individually had a lower scores regarding recall, so we have used them only in combination with the J48 or RF classifiers.

Tables 3 and 4 show the performance of the systems submitted to sub-track 1 and sub-track 2, respectively. In both tables, Training, Development and Test results are presented. Regarding sub-track 1, adding Pat to the classifier seems to improve recall a little, but precision worsens accordingly. Regex does not seem to have hardly any effect. As for sub-track 2, the J48 classifier yields a slightly better precision and slightly worse recall than RF; in both cases, Regex improves recall by 1-3 points but worsens precision by more.

Overall, there are no big differences between the systems submitted, and there is a clear drop in recall in the Test dataset for all. The results seem to be competitive, but official results of other participants in the track have not been published at the time of writing, so no remarks can be made in the matter.

5 Concluding Remarks

In this paper we present the results of applying different machine learning approaches combined with heuristics based on pattern matching and regex based

	Training			Development			Test		
Recognition method	P	R	F1	P	R	F1	P	R	F1
J48	94.09	83.56	88.51	96.75	84.18	90.03	88.12	74.79	80.91
J48 + Pat	90.94	87.76	89.32	93.45	88.70	91.01	87.38	75.63	81.08
J48 + Regex	94.09	83.56	88.51	96.75	84.18	90.03	88.56	74.79	81.71
J48 + Regex + Pat	90.61	87.76	89.16	93.45	88.70	91.01	88.29	76.05	81.71

Table 3. Results of sub-track 1 on the BARR2 Training, Development and Test sets

	Training			Development			Test		
Recognition method	P	R	F1	P	R	F1	P	R	F1
J48	91.92	82.81	86.81	90.28	78.83	84.17	87.57	70.20	77.93
RF	90.46	84.20	87.22	89.71	80.06	84.61	86.41	70.44	77.61
J48 + Regex	85.58	85.82	85.70	84.75	82.63	83.67	81.58	73.36	77.25
RF + Regex	85.71	85.63	85.68	85.79	83.60	84.68	81.72	72.89	77.05

Table 4. Results of sub-track 2 on the BARR2 Training, Development and Test sets

on abbreviation dictionaries. The results show that both tasks are similar in terms of precision, recall and F1-measure when seen from the perspective of the presented results. However, the tasks are quite different, being two different problems that only share partially the detection of abbreviations. Sub-track 1 aims for detecting definitions expressed in the text, and sub-track 2 aims for having it in a dictionary. The dictionary has to be precise and sometimes fails due to changes in the language of the abbreviation or spelling mistakes.

Additionally, there were some exceptions or different abbreviations that we did not contemplate because the task description was not telling this such as:

```
S1889-836X2015000100003-1 SHORT_FORM 398 402 P1NP SHORT-LONG LONG_FORM
404 452 propéptido amino-terminal del procolágeno tipo 1
```

related to:

```
...resultado en los niveles del P1NP (propéptido amino-terminal del
procolágeno tipo 1)...
```

which to our first understanding was not at all the goal of sub-track1, which had to be in the other way round.

Overall, we present a robust method for detecting abbreviations in two different scenarios showing similar results.

6 Acknowledgments

This work has been supported by Vicomtech and the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) under the project TUNER (TIN2015-65308-C5-1-R).

References

1. Intxaurrondo, A., Marimon, M., Gonzalez-Agirre, A., Lopez-Martin, J.A., Rodriguez Betanco, H., Santamaría, J., Villegas, M., Krallinger, M.: Finding mentions of abbreviations and their definitions in Spanish Clinical Cases: the BARR2 shared task evaluation results. In: SEPLN 2018 (2018)
2. Intxaurrondo, A., de la Torre, J.C., Rodriguez Betanco, H., Marimon, M., Lopez-Martin, J.A., Gonzalez-Agirre, A., Santamaría, J., Villegas, M., Krallinger, M.: Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus. In: SEPLN 2018 (2018)
3. Laguna, J.Y.: Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias (2003)
4. Markov, Z., Russell, I.: An introduction to the weka data mining system. ACM SIGCSE Bulletin **38**(3), 367–368 (2006)
5. Montoya, I.: Análisis, normalización, enriquecimiento y codificación de historia clínica electrónica (HCE). Master's thesis, Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra, Euskal Herriko Unibertsitatea (UPV/EHU) (2017)