Finding mentions of abbreviations and their definitions in Spanish Clinical Cases: the BARR2 shared task evaluation results

Ander Intxaurrondo[1,2], Montserrat Marimon[2], Aitor Gonzalez-Agirre[1,2], Jose Antonio Lopez-Martin[3], Heidy Rodriguez[1], Jesus Santamaria[1], Marta Villegas[1,2], and Martin Krallinger[1,2]

[1] Centro Nacional de Investigaciones Oncologicas (CNIO)
[2] Barcelona Supercomputing Center (BSC)
[3] Hospital 12 de Octubre - Madrid

Abstract. A common characteristic of content generated by healthcare professionals, regardless the actual clinical discipline or language, is the widespread and frequent use of abbreviations, acronyms, telegraphic phrases and shorthand notes. Despite the well-known issues related to the ambiguity and misinterpretation of abbreviations, their use in practice is required to simplify and enable communication-avoiding repetition of long complex specialized medical terminologies. Moreover, clinical texts typically do not provide explicit abbreviation definitions. Thus the performance of clinical natural language processing and text mining systems is significantly affected by the previous recognition and definition resolution of medical abbreviations. To promote the development of such key components, we have organized the second Biomedical Abbreviation Recognition and Resolution (BARR2) track. The overall aim of this effort was to evaluate strategies for detecting automatically mentions of abbreviations in running text, as well as returning their corresponding definition given the corresponding context from Spanish clinical case studies. For this track, we constructed the *Spanish clinical case corpus (SPACCC)*. This collection was exhaustively annotated by hand by domain experts with abbreviation mentions together with their corresponding definitions, resulting in the *BARR2 corpus*. A total of 5 teams submitted 26 runs for the two BARR2 subtasks: (a) the detection of explicit occurrences of abbreviation-definition pairs and (b) the resolution of abbreviations regardless whether their definition is mentioned within the actual document. Here we summarize the BARR2 track setting, the obtained results and the methodologies used by participating systems. The BARR2 task summary, resources and evaluation tool for testing systems beyond this campaign are available at: `http://temu.bsc.es/BARR2`.

Keywords: Abbreviation Recognition · Clinical Case Studies · Natural Language Processing

1 Introduction

The problem of finding the correct definition of an ambiguous medical abbreviation can be regarded as a Word Sense Disambiguation (WSD) task where the different definitions are the senses of the medical abbreviation. Abbreviations are also widely used beyond the medical or scientific field. For instance on the AcronymFinder.com website, one of the largest currently available resources of acronyms, an average of 37 new human-edited acronym definitions are added every day [12].

Correct interpretation of abbreviations is not only relevant for automated text-processing systems, but also even for medical professionals themselves and may result in patient safety issues. For instance the abbreviation MTX can easily be misinterpreted as *mitoxantrona* instead of *metotrexato*. A study by Das-Purkayastha et al, even used questionnaire to specifically evaluate how well abbreviations were understood by junior doctors [6]. Using unsuitable abbreviations in prescriptions can cause medication errors [18], and error-prone or other unapproved abbreviations are in fact frequently used in hospitals [18]. Difficulties in interpretation of abbreviations by healthcare professional was observed and resulted in the proposal of standarised abbreviations to circumvent misunderstanding [21] or to spell out abbreviations [3]. Even the use of forced correction alerts of abbreviations had been explored to lower medication errors [16]. Recommendations related to the use of abbreviations for medication or to express dose, route and frequency of administration were made by the Instituto para el Uso Seguro de los Medicamentos [13]. Handling of abbreviations has also been examined for other scenarios, for instance in exercises related to translations of medical texts [24], while it is clear that abbreviation expansion is key for question interpreter modules used by question answering systems [7].

Medical abbreviation recognition and resolution has been studied extensively for English. For instance word embedding based models for acronym disambiguation have been analyzed by Li et al. [12], while others tried to disambiguate context using syntactic features and bag-of-words [8] or topic models [25]. Kim and colleagues tested learning-to-rank models to rank candidate definitions together with external resources like MEDLINE and Unified Medical Language System (UMLS) [11]. An effort to benchmark five teams providing systems for medical abbreviation detection for English was carried out at the ShARe/CLEF eHealth evaluation lab 2013 using strict accuracy and relaxed accuracy measures [22, 17]. Only few annotated resources are available for Spanish, such as the Spanish Radiology Report corpus which also covered the annotation of non-standard abbreviations for a particular clinical discipline and document type [4].

The second Biomedical Abbreviation Recognition and Resolution (BARR2) 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 clinical cases written in Spanish. This task is the follow up of the first BARR track[4].

This paper describes the BARR2 track, as well as the results obtained for this track. Section 2 describes the track setting and posed tasks. Section 3 provides a sort summary of the corpus and resources provided for the track. In section 4 we give a brief explanation of the used evaluation measures. Section 5 provides an overview of the obtained results. Finally, section 6 offers concluding remarks.

2 Task Description

The BARR2 track was posed at the IberEval 2018 evaluation campaign, which had the aim of promoting the development of language technologies for Iberian languages. The purpose of the BARR2 track was to explore settings that are relevant for processing both medical texts and clinical research narratives. The underlying assumption here was that techniques tailored to medical literature could be potentially adapted for processing clinical texts [14].

In essence, the track evaluated the performance of systems for detecting abbreviation-definition pairs in Spanish clinical cases studies and to resolve abbreviations mentioned in text regardless whether its corresponding definition was mentioned in same document. 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 BARR2 track was divided into two separate tasks, which were carried out on the same datasets. The first task focused on the detection of the actual pairs of SF/LF mentions in running text. The second, and main task, focused on the detection of abbreviation mentions in terms of their corresponding character offsets together with the resolution of their corresponding abbreviation definitions.

Participating systems were provided with a training set to construct their predictor or system during the training phase. 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 [9]. 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 BARR2 track we used a particular setting related to the test set release, similar to the evaluation scenario used for BARR track at IberEval 2017 [10]. 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

[4] IberEval 2017. `http://temu.bsc.es/BARR`

of results prior to submission would be unfeasible. Each participating team was allowed to submit for each of the tasks a total of up to five predictions (runs).

3 Data sets

The BARR2 corpus was created after collecting 3,343 clinical cases from SciELO[5] (Scientific Electronic Library Online), an electronic library that gathers electronic publications of complete full text articles from scientific journals of Latin America, South Africa and Spain. A clinician classified those cases into those that were similar to real clinical texts in terms of structure and content and those that were not suitable for this task. Figure legends were automatically removed and in case multiple clinical cases were listed, these were split into single clinical cases.

From these reports, 318 were selected for the training set, 146 for the development set, and 220 for the testing set. These reports were manually annotated by domain experts, using a customized version of Annotator[6] and the Brat[7] annotation toolkit to manually revise mention annotations and annotate the relations between abbreviations and definitions co-mentioned in sentences.

The selected clinical cases are available in *txt* format, encoded with UTF-8. We also made available a file in tabular format with the main information about each clinical case; this file includes the case identifier in the article, ISSN code of the journal, publication date, name of the journal, and a link to the complete full text of the article at the SciELO website. Examining the actual clinical disciplines represented by the BARR2 corpus showed that it covered a broad range of key medical fields, including ophthalmology, urology, digestive diseases, surgery, primary care, pediatrics, internal medicine, nephrology, plastic surgery, intensive care, pharmacy, and oncology. The manual labeling of abbreviation mentions of the corpus was done using a customized version of Annotator. Then, the Brat annotation toolkit was used to revise manually mention annotations and to annotate the relations between short forms and their corresponding long forms, as well as nested mentions.

We refer the reader to the additional material working note paper [9] for a detailed description of the corpus and its annotation process, statistics and annotation consistency analysis.

4 Evaluation

We developed an evaluation script that supported the evaluation of the predictions of the participating teams. The primary evaluation metric used for the BARR2 track consisted in micro-average F-measure. This script provided micro-average standard performance statistics, such as precision, recall, and F-score,

[5] `http://www.scielo.org`
[6] `https://github.com/openannotation/annotator/`
[7] `http://brat.nlplab.org/`

and enabled the examination of annotation mismatches. The source code of the script is available at the track's website.

During the test phase, teams were requested to generate predictions for a blinded collection of documents and they had to send their submission to the organisers within a short period of time. Teams could submit up to five prediction files (runs). For the first task, the evaluation was very strict. Besides mentioning the abbreviation-definition pairs, participants had to specify the exact positions of the mentions in each document, returning the starting and ending offsets.

The evaluation of the second task had a considerable complexity when evaluating the definitions given by participants. We must take into account that an abbreviation may have a single definition, but there could be many variants for that definition, including typographical variants or other aliases. For example, the abbreviation *IV* may have the definition *"intravenoso"*, and participant may have predicted it as *"intravenosa"*. Both definitions are very similar and both are correct, since they only differ in the gender. To make the evaluation easier, we lemmatized the manually annotated definitions, and added them to the gold standard, together with the originals. Participants were also asked to lemmatize their definitions and to provide both versions, so that we could compare both of them and avoid correct definitions being considered incorrect. We used the IXA-pipes pipeline [1] for the lemmatization process, participants could make use of their prefered lemmatizers.

The evaluation of the second task admitted predicted definition tokens present in gold definitions, giving scores between 0 and 1. To calculate the score, we separated each token of the gold and predicted definitions, detected the stop words using a list[8] and removed them, and later we checked the number of tokens in the predicted definition that matched with the ones in the gold definition.

The following algorithm summarizes the evaluation process:

1. Check if the found abbreviation is mentioned at the gold annotations and offsets match. If correct, go to the next step. If wrong, return 0.
2. Check if the guessed definition matches with gold definition, if so, return 1. If not, check if the lemmatized definition matches with the gold lemmatized definition. If correct, return 1. If not, go to the next step.
3. Tokenize predicted and gold definition and remove stop words. Check if the number of tokens, and the tokens themselves, match. If so, return 1. If not, repeat this process with lemmatized definitions. If they do not match, go to the next step.
4. Check if the tokens at the predicted definition are present in the gold definition. Divide the number of tokens present with the number of maximun tokens between gold annotation and the definition. Repeat this process with the lemmatized definitions. Check both divisions, return the highest score.

The evaluator displays three different evaluation results:

[8] `https://github.com/stopwords-iso/stopwords-es/blob/master/stopwords-es.txt`

BARR2 sub-track 1	Run ID	Precision	Recall	F1-score
Fsanchez	1	88.61	**88.23**	**88.42**
Vicomtech	regex+ML+pat	88.29	76.05	81.71
Vicomtech	ML+regex	88.56	74.79	81.09
Vicomtech	ML+pat	87.34	75.63	81.08
Vicomtech	ML	88.12	74.79	80.91
UNED	3	85.36	73.53	79.01
UNED	4	83.98	72.69	77.93
UNED	5	84.84	70.59	77.06
UNED	1	85.13	69.75	76.67
UNED	2	**91.20**	47.90	62.81

Table 1. Results of the evaluation of abbreviation-definitions pairs sub-track.

- Ultra-strict evaluation: the evaluator considers correct only those definitions that exactly match with the gold annotations. Returns 0 if wrong at steps 2 (from the list above).
- Strict evaluation: the evaluator considers correct only those definitions where all tokens present at the definition match with all tokens at the gold definition, the order does not matter. Returns 0 if wrong at step 3.
- Flexible evaluation: step 4.

For this sub-track, flexible evaluation results were considered as official, leaving the results of the other two evaluation types for error analysis.

5 Participants, Results and Discussion

A total of 5 teams participated in the track. We refer the reader to participants' papers ([20, 2, 19, 15, 5]) for a full description of the systems they developed. In this section, we present the results they achieved and an error analysis. The evaluated BARR2 teams submitted a total of 26 runs: 9 for the first task and 17 for the second one.

Table 1 illustrates the obtained performance for each of the evaluated submissions for the first task, focused on the detection of pairs of SF/LF mentions in running text. Top-scoring team was Fsanchez, with an F-score of 88.42%, followed by Vicomtech's highest score 7 points below and UNED's highest score 9 points below.

Here we summarize different errors found with some examples:

- Short form in English, definition written in Spanish:
  - *SIADH → Syndrome of Inappropiate Secretion of Antidiuretic Hormone* (*Síndrome de Secreción Inadecuada de Hormona Diurética*).
- Nested forms mistaken as long forms:
  - *(...) Ecocardiografía Transtorácica (ETT) y Transesofágica (ETE) (...) → Ecocardiografía Transtorácica* = long form.
- Abbreviation near parenthesis, but definition not present nearby:
  - *(...) los dominios 'TOM' (número de errores no-significativo, aunque (...) → TOM = Teoria de la Mente.*
- Compound abbreviations with complex definitions:
  - *ECO-PAAF = Punción-Aspiración con Aguja Fina Guiada por Ecografía*

BARR2 sub-track 2	Run ID	Precision	Recall	F1-score	F1 strict	F1 canonical
Fsanchez	1	85.34	**81.11**	**83.17**	79.85	82.89
Hospital-italiano	3ul	**88.90**	71.29	79.13	77.36	79.67
Hospital-italiano	4ul	87.08	71.64	78.61	76.80	79.73
Hospital-italiano	4ul-sinpl	86.95	71.54	78.50	76.71	79.73
Vicomtech	ML	87.57	70.20	77.93	75.65	79.30
Vicomtech	MLRF	86.41	70.44	77.61	75.51	79.09
Vicomtech	ML+regex	81.58	73.36	77.25	74.88	78.80
Vicomtech	MLRF+regex	81.72	72.89	77.05	74.68	78.58
Hospital-italiano	3ul-hiba	83.77	69.39	75.90	73.20	75.19
Hospital-italiano	4ul-hiba	75.97	67.85	71.68	68.87	71.91
UC3M	2	74.93	37.69	50.16	46.82	50.72
UC3M	3	74.92	37.69	50.15	46.82	50.72
UNED	3	41.80	24.24	30.69	28.15	29.12
UC3M	1	59.80	19.43	29.33	26.18	30.43
UNED	1	38.88	22.54	28.54	25.81	27.04
UNED	2	36.92	21.41	27.10	24.78	26.44

Table 2. Results of the evaluation of abbreviation recognition and resolution sub-track.

Table 2 illustrates the obtained performance for each of the evaluated submissions for the second, and main task, focused on the recognition and resolution of abbreviations. Fsanchez became the top-scoring team for this task as well, with an F-score of 83.17%, followed by Hospital-Italiano, with an F-score of 79.13%, and Vicomtech with 77.93%. The last column displays the results obtained by participants by using normalized definitions present in SNOMED CT to evaluate participants' definitions, this evaluation process was still preliminary.

Here we summarize different errors found with some examples:

- Abbreviations ending with dots, when the dot indicated the end of the sentence and not the abbreviation:
  - *cm., comp.*, etc.
- Abbreviations with just a single character, but many possible meanings:
  - *N* = *Neutrófilo* or *Nitrógeno*
  - *L* = *Listeria* or *Litro*.
- Short forms that looked like Roman numbers:
  - *IV* = *intravenosa* or *4*
  - *VI* = *Ventrículo Izquierdo* or *6*.
- Abbreviations with many definitions:
  - *FA* = *Fosfatasa Alcalina* or *Fibrilación Auricular*
  - *RM* = *Regurgitación Mitral* or *Resonancia Magnética*
  - *DC* = *Donante Cadavérico* or *Diagnóstico Clínico*
- Definitions' variants, same meaning but written in different forms:
  - *TAC* = *Tomografía Axial Computarizada* or *Tomografía Axial Computerizada*
  - *LDH* = *Lactato Deshidrogenosa* or *Lactato-deshidrogenasa*

6 Conclusions

The BARR2 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. This 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 [23]. 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.

It is important to highlight the complexity of the task, as some of the abbreviations and even the definitions corresponded to terms in Spanish, others to terms in English. On the other hand the issue of alternative correct definition variants should be addressed in future follow up studies. We are exploring the used of concept normalizations of definitions to existing knowledge-bases including SNOMED CT, UMLS, MeSH and UniProt as an alternative strategy to standardize abbreviation mentions.

Examining which are the most frequent types of abbreviations, we observed that many of them corresponded to units of measure (key for detection of posology and dosage), anatomical entities, biochemical markers and treatments. When looking at the language of the definitions of abbreviations, only around 68 percent corresponded to Spanish definitions, the remaining where mostly English definitions often related to substances, treatments and biochemical entities. We manually mapped 500 definitions to SNOMED. Examining the corresponding concept class showed that most corresponded to the class substance.

Acknowledgements

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

References

1. Agerri, R., Bermudez, J., Rigau, G.: Ixa pipeline: Efficient and ready to use multilingual nlp tools. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
2. Castaño, J., Ávila, P., Pérez, D., Berinsky, H., Park, H., Gambarte, L., Luna, D.: A simple approach to abbreviation resolution at barr2, ibereval 2018. SEPLN (2018)
3. Cheng, T.O.: Medical abbreviations. Journal of the Royal Society of Medicine **97**(11), 556–556 (2004)
4. Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of entities and relations in spanish radiology reports. In: RANLP. pp. 177–184 (2017)
5. Cuadros, M., Pérez, N., Montoya, I., García-Pablos, A.: Vicomtech at barr2: Detecting biomedical abbreviations with ml methods and dictionary-based heuristics. SEPLN (2018)
6. Das-Purkayastha, P., McLeod, K., Canter, R.: Specialist medical abbreviations as a foreign language. Journal of the Royal Society of Medicine **97**(9), 456–456 (2004)
7. Diekema, A.R., Yilmazel, O., Chen, J., Harwell, S., He, L., Liddy, E.D.: Finding answers to complex questions (2004)

8. HaCohen-Kerner, Y., Kass, A., Peretz, A.: Combined one sense disambiguation of abbreviations. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. pp. 61–64. Association for Computational Linguistics (2008)
9. Intxaurrondo, A., de la Torre, J., Rodriguez, H., Marimon, M., Lopez-Martin, J., 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. BARR2 track additional material of SEPLN paper (2018), `http://temu.bsc.es/BARR2/downloads/corpus-BARR2.pdf`
10. Intxaurrondo, A., Pérez-Pérez, M., Pérez-Rodríguez, G., López-Martín, J.A., Santamaria, J., de la Pena, S., Villegas, M., Akhondi, S.A., Valencia, A., Lourenço, A., Kralllinger, M.: The biomedical abbreviation recognition and resolution (barr) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts (2017)
11. Kim, J., Oh, H., Nam, S., Myaeng, S.H.: A method using candidate exploration and ranking for abbreviation resolution in clinical documents. In: 2013 IEEE International Conference on Healthcare Informatics (ICHI). pp. 317–326. IEEE (2013)
12. Li, C., Ji, L., Yan, J.: Acronym disambiguation using word embedding. In: AAAI. pp. 4178–4179 (2015)
13. López, M.O., Muñoz, R.M., Hurlé, A.D.G.: Seguridad de medicamentos abreviaturas, símbolos y expresiones de dosis asociados a errores de medicación. Farmacia Hospitalaria **2004**(28/2), 141 (2004)
14. 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)
15. Montalvo, S., Martínez, R., Almagro, M.: Mamtra-med at biomedical abbreviation recognition and resolution - ibereval 2018. SEPLN (2018)
16. Myers, J.S., Gojraty, S., Yang, W., Linsky, A., Airan-Javia, S., Polomano, R.C.: A randomized-controlled trial of computerized alerts to reduce unapproved medication abbreviation use. Journal of the American Medical Informatics Association **18**(1), 17–23 (2010)
17. Patrick, J.D., Safari, L., Ou, Y.: Share/clef ehealth 2013 normalization of acronyms/abbreviations challenge. In: CLEF (Working Notes). Citeseer (2013)
18. Samaranayake, N., Dabare, P., Wanigatunge, C., Cheung, B.: The pattern of abbreviation use in prescriptions: a way forward in eliminating error-prone abbreviations and standardisation of prescriptions. Current drug safety **9**(1), 34–42 (2014)
19. Sánchez, C., Martínez, P.: A simple method to extract abbreviations within a document using regular expressions. SEPLN (2018)
20. Sánchez-León, F.: Arborex: Abbreviation resolution based on regular expressions for barr2. SEPLN (2018)
21. Sheppard, J.E., Weidner, L.C., Zakai, S., Fountain-Polley, S., Williams, J.: Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. Archives of disease in childhood (2007)
22. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., et al.: Overview of the share/clef ehealth evaluation lab 2013. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 212–231. Springer (2013)
23. 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)

24. Ynfiesta, B.B., Suárez, L.T., Peraza, A.V.F.: Translation of acronyms and initialisms in medical texts on cardiology. CorSalud (Revista de Enfermedades Cardiovasculares) **5**(1), 93–100 (2013)
25. Zhang, W., Sim, Y.C., Su, J., Tan, C.L.: Entity linking with effective acronym expansion, instance selection, and topic modeling. In: IJCAI. vol. 2011, pp. 1909–1914 (2011)