Resources, guidelines and annotations for the recognition, definition resolution and concept normalization of Spanish clinical abbreviations: the BARR2 corpus

Ander Intxaurrondo[1,2], Juan Carlos de la Torre, Heidy Rodriguez[1], Montserrat Marimon[2], Jose Antonio Lopez-Martin[3], Aitor Gonzalez-Agirre[1,2], 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. One of the main bottlenecks for clinical natural language processing is the lack of access to annotated corpora generated by domain experts. Due to the complexity of the medical language and the heavy use of specialized terminologies the construction of labor-intensive Gold Standard datasets within the biomedical and clinical field is of critical importance and practical value. In order to increase the competitive use of manually annotated clinical case studies, a specialized type of medical article that shows the highest degree of similarity to real clinical texts, we have constructed the second Biomedical Abbreviation Recognition and Resolution (BARR2) corpus as part of a shared task posed at IberEval 2018. This corpus contains Spanish clinical case study sections from a variety of clinical disciplines. The BARR2 clinical cases were annotated by hand in an exhaustive manner by domain experts to label not only all abbreviation mentions but also their corresponding definitions. A subset of these definitions was also manually mapped to control vocabulary resources, in particular to SNOMED CT. We have used this corpus as a resource to train and evaluate participating systems of the BARR2 track. Here we summarize the creation of the BARR2 corpus and the annotation process to create the training, testing and development set for participants. More information about this campaign can be found at: `http://temu.bsc.es/BARR2`.

Keywords: Abbreviation · Clinical Case · Natural Language Processing · Corpus

1 Introduction

Clinical texts do contain a considerable fraction of abbreviation mentions, as they constitute a way to provide a more compact representation of medical expressions [2]. To interpret correctly some highly ambiguous abbreviations is cumbersome for health care practitioners and can be a cause of medical errors

[12, 5]. Acronyms in case of English biomedical texts are overloaded 33% of the time [8]. This also implies that manual annotation and resolution of their definition is a highly time consuming task which requires domain expertise and often consultation of external medical resources.

Clinical narratives normally do not show explicit co-mentions of abbreviations (short forms) together with their corresponding full versions, descriptive forms or definitions (long forms), also known as abbreviation-definition pairs. In order to be able to resolve abbreviation definitions and to construct sense inventories the access to resources covering abbreviation definitions is important. While many resources like ALLIE, UMLS or Acromine exist for English medical abbreviations, only few resources were generated for Spanish that might be serve to interpret abbreviations. Navarro described a resource for abbreviations and acronyms commonly used in Spanish medical texts [9].

Finding and resolving medical abbreviations was intensively examined for English [13], whereas limited efforts were made for Spanish medical abbreviation recognition.

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 BARR track of IberEval 2017 tried to promote the construction of abbreviation definition resources from the Spanish medical literature [7]. Following up the interest in the first BARR track we tried to promote not only the extraction of explicitly co-mentioned abbreviation-definition pairs but to encourage the implementation of real world abbreviation resolution systems regardless of the co-occurrence of abbreviation definitions within the same document.

Therefore the second Biomedical Abbreviation Recognition and Resolution (BARR2) track provided Gold Standard training, development and test corpora manually annotated by domain experts with abbreviation-definition pairs within clinical cases written in Spanish. We refer the reader to [6], where we focus on the description of the track, the evaluation measures, and participants' results.

2 BARR2 Corpus

The main barrier for the development of Spanish clinical text processing tools, is the difficulty of accessing electronic health records. Patient privacy and reveal of unfavorable institutional practices are enough reasons for hospitals and clinics to be extremely reluctant when allowing access to clinical data for researchers from outside the associated institutions [3]. Therefore the use of clinical literature or synthetic surrogate corpora was explored, not only for Spanish but also in case of English or German clinical NLP.

The sources we used to obtain open access case reports included PubMed, IBECS and SciELO:

– PubMed[4] is a free search engine used to access Medline database, a bibliographical database of references and abstracts on life sciences and biomedical

[4] `https://www.ncbi.nlm.nih.gov/pubmed/`

topics. It is maintained by the U.S. National Library of Medicine. It contains more than 28 million publications in 39 languages, including Spanish.

- IBECS[5] (Spanish Bibliographic Index in Health Sciences) is a bibliographic database, maintained by the Spanish National Health Sciences Library of the Carlos III Health Institute. It collects scientific journals covering multiple fields in health sciences published in Spain.
- SciELO[6] (Scientific Electronic Library Online) is an electronic library supported by the Sao Paulo Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (BIREME). This initiative gathers electronic publications of complete full text articles from scientific journals of Latin America, South Africa, and Spain.

We can find several cases with links to free full text available at PubMed and IBECS. PubMed contains direct links to 272,647 open access case reports present in external websites, from which 242,930 are in English and 11,654 in Spanish. Meanwhile, IBECS contains direct links to 31,998 clinical cases in Spanish. Some links from PubMed and IBECS point us to different reports present at the SciELO network.

In order to provide a document collection of Spanish clinical cases for the BARR2 track, we gathered the articles from SciELO and automatically extracted the sections related to clinical cases from the articles using several patterns; the articles where we were unable to detect the section were discarded. The final BARR2 document or background collection contains a total of 3,343 records; 69 belong to SciELO Argentina, 99 to SciELO Brazil, 744 to SciELO Chile, 65 to SciELO Columbia, and, finally, 2,490 to SciELO Spain. Each report is identified with the same identifier used by SciELO, and may include a single clinical case or more. We separated each clinical case following different patterns. The final identifier for each case is formed by the SciELO code and the number of case inside the article. Let's consider the following report ID: *S1137-66272011000100013-3*; for this ID, the first and second sequences together (*S1137-66272011000100013*) form the SciELO identifier, while the last number (*-3*) indicates that this case is the third one found in the SciELO article.

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.

[5] `http://ibecs.isciii.es`
[6] `http://www.scielo.org`

	Train	Development	Background	Test
Total number of reports	318	146	2,879	220
Total number of sentences	4,781	2,262	55,008	3,812
Maximum number of sentences	52	66	209	64
Minimum number of sentences	3	4	1	4
Mean number of sentences	15.03	15.49	19.11	17.33
Total number of tokens	122,594	56,564	1,264,527	90,098
Maximum number of tokens	1,063	996	5,678	1,280
Minimum number of tokens	76	84	27	100
Mean number of tokens	385.52	387.42	439.22	409.54
Total number of sentences in complete corpus: 62,051				
Total number of tokens in complete corpus: 1,433,685				

Table 1. BARR2 corpus statistics.

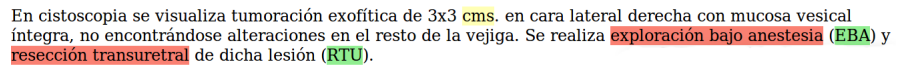

Fig. 1. Annotator screenshot for clinical case *S0004-06142008000100008-1* (development set).

3 Annotation

A subgroup from the BARR2 document collection was used to construct a manually labelled training and test set. We cautiously selected 684 clinical cases from SciELO Spain, which we distributed as follows: 318 clinical cases for the training set, 146 for the development set and 220 for the test set. Table 1 provides a statistical overview of these datasets, covering basic corpus statistics.

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. Annotator[7] is a library for text annotation in browsers which provides a set of interoperable tools for annotating content in webpages. Brat[8] [11] is another intuitive web-based tool for text annotation for a variety of NLP tasks. Unlike Annotator, Brat gives the possibility to annotate relations between different entities.

Fig. 1 shows an screenshot of the Annotator interface. The annotating process was simple: annotators selected the mentions they wanted to annotate, this showed a pop-up window with the possible labels, and they chose the desired label. For each abbreviation labeled, annotators had to specify their definitions as well; at the *"Comments"* section, they could write the abbreviation's meaning.

Fig. 2 shows an screenshot of the Brat interface. To label relations, annotators selected short forms (abbreviations) and dragged the pointer to the corresponding long forms (definitions). For nested relations, they pointed abbreviations with nested cases, and these with definitions (Fig. 3).

[7] `https://github.com/openannotation/annotator/`
[8] `http://brat.nlplab.org/`

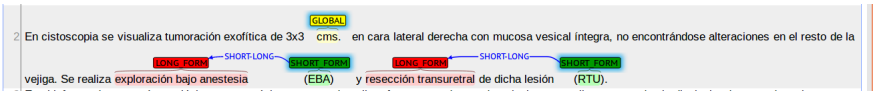

Fig. 2. Brat screenshot for clinical case *S0004-06142008000100008-1* (development set).

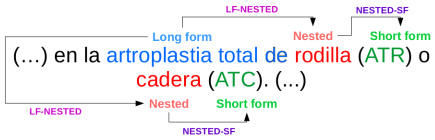

Fig. 3. Nested relation example.

The labels used to annotate abbreviations were the same used for the first BARR track. We used *Short-Form, Long-Form and Nested* for the first subtrack and *Nested, Multiple, Global, Contextual, Unclear and Derived* for the second one[9].

In order to make the annotation process as flexible as possible, annotators had access to the complete article of the clinical case to find their meanings explicitly mentioned in other sections. They also used several Spanish biomedical abbreviation and acronym dictionaries. In addition, we applied an automatic abbreviation-definition pairs detection algorithm [10] to detect different pairs in the full text, so annotators did not need to read all the article looking for explicit definitions, unless it was necessary. We also used an internal Spanish abbreviation-definition pair database[10] to assign definitions to those abbreviations with no explicit definitions.

Table 2 shows the frequency of mention types and relation types in the training set, development set, and testing set.

3.1 Quality checking

During the transition from Annotator to Brat, we applied several postprocessing algorithms to improve the quality of the corpus. Postprocessing included the following steps:

1. Automatic annotation of unannotated definitions for frequent abbreviations.
2. Automatic detection of unfrequent abbreviations without definitions.
3. Automatic detection of unannotated abbreviations, previously annotated in the corpus.

All these changes could be seen directly at Brat.

[9] In fact, in the second subtrack, the name of the label was irrelevant.
[10] `https://github.com/Med-TL/Plan-TL/tree/master/AllieES`

	Train	Development	Test
Total number of reports	318	146	220
Subtrack 1			
Mention type	**Train**	**Development**	**Test**
LONG_FORM	280	177	232
SHORT_FORM	280	177	232
NESTED	12	0	12
Relation type	**Train**	**Development**	**Test**
SHORT-LONG	274	177	266
SHORT-NESTED	6	0	6
NESTED-LONG	6	0	6
Subtrack 2			
	Train	**Development**	**Test**
Abbreviations	4,260	1,878	3,414

Table 2. Manually annotated abbreviations and relations in the BARR2 corpus.

In the first step, in order to speed the annotation of the corpus, we asked the annotators not to provide the definitions to highly frequent abbreviations. Instead, we automatically annotated them with the definitions they had been assigned previously in the corpus.

This algorithm also allowed us to detect those annotated abbreviations without definition which had not been annotated previously. In this situation, the annotators were asked to provide a definition in the Annotator version of the clinical case.

Finally, we detected unannotated abbreviations. In this step, we used the list of previously annotated abbreviations and looked for them in clinical cases. If the algorithm detected an unannotated token that matched an abbreviation at this list, we annotated it automatically and assigned the definition as well. Annotators could check and decide if the automatically detected abbreviation and its definition were correct or not.

3.2 Distributed files

Due to the complexity to evaluate the definitions for subtrack 2, we lemmatized them and included them in the annotation files. We used the IXA-pipes pipeline [1] for the lemmatization process, participants could make use of their prefered lemmatizers.

Training, development and testing sets were distributed each in two different files: annotations for the first subtrack, and the annotations for the second subtrack. These files were in tabular format (*TSV* extension), and used the UTF-8 encoding. For each abbreviation found in text, these files included their starting and ending offsets, we did the same for explicit definitions annotated in text for subtrack 1. For subtrack 2, we added the originally annotated definition and its lemmatized version at the end of the line.

3.3 Inter-annotator agreement

We evaluated the inter-annotator agreement (IAA) between the two different annotators that worked on the manual annotation process. For this analysis,

Total annotations	879	100%
Annotations made by 1	854	97%
Annotations made by 2	826	93%
Mention annotation coincidence	851	96%
Mention annotation coincidence (same label)	781	88%

Table 3. Inter-annotator agreement analysis results in numbers.

we randomly selected 34 clinical reports from the training set, and 16 from the development set, having a total of 50 clinical cases.

To calculate the IAA, we obtained the total number of annotations made by both annotators, and the number of annotations made by each annotator. We checked if the annotations made by both matched the starting and ending offsets, and if they did, then we also checked if the labels they assigned to each annotated mentions matched.

Table 3 shows the results of the IAA analysis. Results show there are very little discrepancies between both annotators. Differences between labels occur more often, but they are irrelevant for the task: annotators used the same labels of the first edition of BARR, which had different ways to annotate abbreviations; for example, one annotator could have annotated a certain short form as *Global* and the other annotator as *Contextual*. Labels may be different, but at least abbreviations were correctly annotated. Situations where a long form was annotated as a long form, or the opposite case, never happened.

Analyzing the IAA for the manually annotated abbreviations' definition was left out, because it could not have been done properly. As we mentioned above, annotators only specified the definition of very common non-ambiguous abbreviations once or twice, and those abbreviations without definition got their definition assigned automatically.

4 Conclusion and discussion

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 [14].

The corpus created for the track was composed of different open access clinical cases written in Spanish. Domain experts annotated abbreviations present in these reports, together with their definition. Inter-annotator agreement results showed that annotators agreed on the 88% of the labeled cases, concluding that the guidelines they used were clearly structured.

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 En-

glish 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. Benavent, R.A., Iscla, A.A.: Problemas del lenguaje médico actual.(ii) abreviaciones y epónimos. Papeles Méd **10**(4), 170–6 (2001)
3. Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K., Uzuner, O.: Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association **18**(5), 540–543 (2011). https://doi.org/10.1136/amiajnl-2011-000465, `http://dx.doi.org/10.1136/amiajnl-2011-000465`
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. on Accreditation of Healthcare Organizations, J.C., et al.: Medication errors related to potentially dangerous abbreviations. Sentinel Event Alert **23**, 1–4 (2001)
6. Intxaurrondo, A., Marimon, M., Gonzalez-Agirre, A., Lopez-Martin, J., Rodriguez, H., Santamaría, J., Villegas, M., Krallinger, M.: Finding mentions of abbreviations and their definitions in spanish clinical cases: the barr2 shared task evaluation results. SEPLN (2018)
7. 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)
8. 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)
9. Navarro, F.A.: Repertorio de siglas, acrónimos, abreviaturas y símbolos utilizados en los textos médicos en español. Panace **9**(27) (2008)
10. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E. (eds.) Pacific Symposium on Biocomputing. pp. 451–462 (2003), `http://dblp.uni-trier.de/db/conf/psb/psb2003.html#SchwartzH03`

11. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: A web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2380921.2380942
12. Tena Marsà, X.: El culto a las abreviaciones: idolatría o virtud. Reumatología Clínica **8**(2), 54–55 (2012)
13. 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)
14. 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)