# A Simple Approach to Abbreviation Resolution at BARR2, IberEval 2018

José Castaño, Pilar Ávila, David Pérez, Hernán Berinsky, Hee Park, Laura Gambarte, Daniel Luna

Departamento de Informática en Salud, Hospital Italiano de Buenos Aires

**Abstract.** Acronyms and abbreviations are widely used in clinical and other specialized texts. Understanding their meaning constitutes an important problem in the automatic extraction and mining of information from text. Moreover, an even harder problem is its sense disambiguation; that is, where a single acronym refers to many different meanings in different texts, a common occurrence in the clinical texts. In such cases, it is necessary to identify the correct corresponding sense for the acronym or abbreviation, which is often not directly specified in the text. Here we present an approach to identify acronyms and abbreviations for the BARR2 competition. We use cTAKES [7] as a framework to develop an approach to identify abbreviations and acronyms as part of a lookup entity recognition system and a word sense disambiguation classifier. The results of the BARR2 test set have shown a 79.13 F measure.



## 1 Introduction

The problem of retrieving the meaning of acronyms and abbreviations in clinical text is related to that of entity identification, since it is necessary to know which entity an acronym or abbreviation expression (also called short forms) refers to in a text in order to accurately identify and extract target information.

The problem of automatically determining the meaning of short forms in medical texts is both a critical as well as a difficult one. It is critical because the performance of information retrieval and extraction tasks is significantly degraded when acronym and abbreviation meanings are not properly understood or interpreted. The problem is exacerbated in the medical literature by the widespread use and frequent coinage of novel short forms and new short form meanings. Furthermore, there is wide variance in conventions within the medical communities on forming acronyms from their "long forms". Acronyms and abbreviations were addressed as a problem to solve in the biomedical domain a long time ago (e.g. [6]). A number of different techniques have appeared that determine automatically the meaning of an acronym in free text. Most of these works distin-

guish between "standard" acronyms on the one hand, and abbreviations and aliases on the other.

Clinical terms may be noisy descriptions typed by healthcare professionals in the electronic health record system (EHR). Description terms contain clinical findings, suspected diseases, among other categories of concepts. Descriptions are very short texts presenting high lexical variability containing synonymy, acronyms, abbreviations and typographical errors. Automatic mapping of description terms to normalized descriptions in an interface terminology is a hard task and it is based essentially on string similarity features. In this scenario, abbreviations and acronyms pose a special challenge for several reasons. The Joint Commission International[1] requires that the use abbreviations must be controlled on patient materials and documents to ensure that patients and their families understand the information available in their records [2]. Also, according to the SNOMED CT Editorial Guide, abbreviations are prohibited in fully specified names and synonyms, with specified exceptions.

The organization of BARR and BARR2 initiates a special effort on this topic for Spanish language. Even if there are some compiled resources there are no public available databases for the clinical domain in Spanish language. There is great variability in the use of abbreviations and acronyms and many of them present high degree of ambiguity.

The problem of sense disambiguation is a crucial one in an information retrieval system. A common acronym such as $AA$ has many different meanings, such as:[3]

- abdomen agudo
- alcohólicos anónimos
- amenaza de aborto
- aminoácido
- anemia aplásica
- aorta abdominal
- aorta ascendente
- apendicitis aguda

The Hospital Italiano de Buenos Aires (HIBA) has an interface Spanish vocabulary [4, 2] where each term is mapped via a direct relation or

[1] The international organization that ensures international accreditation and certification of hospitals and other healthcare centers.

[2] https://www.jointcommissioninternational.org/use-of-codes-symbols-and-abbreviations/

[3] These expansions given at the *Diccionario de Siglas Médicas*[1], there are other used such as *aleteo auricular*, very frequent at HIBA data.

using compositional post-coordinated expressions to SNOMED CT as its reference vocabulary. The local interface vocabulary was implemented in 2002 and it was implemented using those description terms typed by the healthcare professionals. The absence of SNOMED CT support of abbreviations and the troubles caused by the use of abbreviations in clinical records brought the need to create a content extension to detect and disambiguate them and still maintain the standard reference language. The HIBA implemented in 2015 a context extension system of abbreviation recognition consisting of 800 unique abbreviations and 200 ambiguous abbreviations. Also the healthcare professional is able to introduce its own expansion form if none of the possible meanings is the intended one. There are also 1200 abbreviations with no standardized expansion form that are available for expansion to be performed by the healthcare professional.

## 2 The BARR2 track challenge

The Second Biomedical Abbreviation Recognition and Resolution (BARR2) track has *the aim to promote the development and evaluation of clinical abbreviation identification systems*. There are two sub-tracks and we chose (due to time limitations and scope focus) to participate in the Subtrack 2, the abbreviation resolution track. In this case the challenge is to identify the acronyms and abbreviations in the text and to provide the corresponding definition or long form.

The BARR2 organization provided a training set consisting of 318 clinical cases that had been published in the clinical literature and the corresponding metadata for original record and the corresponding journal and publication date. A development set consisting of 146 clinical cases was also released, and finally for the challenge participating teams had to submit their predictions for the background set composed of 2879 clinical cases. The test set consisting of 220 clinical cases was released after the participating groups submitted their predictions for the background set.

## 3 Our approach to short form resolution

We decided to use cTAKES as a framework to test acronym and abbreviation resolution algorithms. The BARR2 source text has not been previously tokenized, so different tokenization algorithms have an impact on the system performance. We used cTAKES pipeline facility to test different parameters. We slightly adapted cTAKES sentensifier, tokenizer,

and we used the universal POS tagger [5] model from OpenNLP available at https://cavorite.com/labs/ nlp/opennlp-models-es/ . Acronyms and abbreviations are identified using cTAKES entity recognizer based in a dictionary lookup strategy implemented in the DefaultJCasTermAnnotator class. Therefore at the lookup phase acronyms and abbreviations were identified, and their possible definitions retrieved. The Stanford Column Classifier (a Maximum Entropy model)[3] was used to disambiguate or filter those long forms that were not predicted by the classifier. We built a model for each ambiguous short form, based on the short form, the clinical case text and the long form to be predicted. Three sources of data for the possible acronym and abbreviation expansion: a) HIBA context terminology, b) Diccionario de Siglas Médicas [1] and c) the BARR2 training data.

An initial assessment has shown that it was very difficult to add the data from Diccionario de Siglas Médicas. In particular, it was not easy to normalize those expansions that had the same meaning but had different long forms, i.e. synonym long forms. Therefore we decided to use only those data for which we had training sets for the classifier, HIBA abbreviations, and BARR2 training set. We split the data in training and test to Evaluate the classifier. Table 3 shows the data used.

| Data | HIBA | BARR2 | Total |
|---|---|---|---|
| Training | 222225 | 2558 | 224783 |
| Test | 79117 | 883 | 80000 |
| Total | 301342 | 3441 | 304783 |
| Ambiguous | 368 | 99 | 522 |

**Table 1.** Data Sets used for training the classifier

Table 2 reports Macro and Micro $F_1$ measures reported by the Stanford classifier using different models combining BARR2 and selected HIBA data. It can be seen that HIBA model performs poorly on BARR2 test data and vice versa.

| Model | BARR2 | HIBA | HIBA-BARR2 |
|---|---|---|---|
| HIBA | 0.199/0.269 | 0.97/0.82 | 0.964/0.736 |
| HIBA-BARR2 | 0.628/0.526 | 0.973/0.83 | 0.968/0.771 |
| BARR2 | 0.869/0.894 | 0.447/0.067 | 0.444/0.356 |

**Table 2.** Micro/Macro $F_1$ measures on the three test sets.

We performed a few tests on the training data and we found that it was very difficult to predict correct expansions. In particular there were some cases of spurious ambiguity:

- virus de epstein-barr vs virus de epstein barr
- tomografía axial computadorizada vs tomografía axial computarizada vs tomografía axial computada

Therefore we took a very simple approach. We selected the most frequent expansions for those abbreviations in the BARR2 training set, in other words there were no ambiguous short forms from the training set. Those expansions that were not predicted by the classifier were discarded. This is what we called the *fq3* model and we obtained a good baseline result using this simple strategy. We used also used the same data combined with the HIBA abbreviations and acronyms, and in this case we did not use any filtering. This is our *fq3-HIBA* model. Finally we used the same strategy using both the training and the development set. Our *fq4* and *fq4-HIBA* models.

## 4 Evaluation and Results

BARR2 organization provided the evaluation tool to be used in the training and development sets. The tool provides three measures, *Ultra-strict, Strict and Flexible Evaluations. Ultra-strict* evaluation requires that the exact same expansion string be predicted. *Strict* evaluation does not consider stop words nor word order, a list of stopwords was provided and uses lemmatized forms. The *Flexible* evaluation used a stemmer to compare predictions. We undestand that the *Strict* evaluation provides a closer comparison, given stopwords at the expansion usually do not provide semantic information, and lemmatized forms preserve meaning. Unfortunately we did not have a lemmatizer ready in the pipeline so we did not use lemmatized forms.

| Model | **$F_1$** | | | **Precision** | | | **Recall** | | |
|---|---|---|---|---|---|---|---|---|---|
| | training | dev. | test | training | dev. | test | training | dev. | test |
| fq3 | 85.19 | 73.53. | 79.13 | 88.03 | 84.66 | 88.90 | 82.53 | 64.97 | 71.29 |
| fq4 | 80.72 | 83.03 | 78.61 | 84.70 | 88.15 | 87.08 | 80.72 | 78.47 | 71.64 |
| fq3-hiba | 80.18 | 70.49 | 75.90 | 79.81 | 79.80 | 83.77 | 80.56 | 63.07 | 69.39 |
| fq4-hiba | 75.70 | 73.64 | 71.68 | 72.82 | 76.29 | 75.97 | 78.82 | 71.17 | 67.85 |

**Table 3.** $F_1$ measure, Precision and Recall Strict Evaluation Results

## 5 Conclusions and Future Work

This work was prepared in a very short time, and the approach we used was very simple. It was apparent from the very beginning that the BARR2 and the HIBA data were very different, and it is reflected in the performance when HIBA data is used. One of the difficulties we faced is the need of using a normalization function, based in string similarity, so as to map to a canonical string. We did not have time also to include a lemmatizer in the pipeline, which might improve a little the results.

Our strategy produced better Precision than Recall results, this can be seen as an effect both of the preprocessing pipeline we used, and also on the filtering use of the classifier.

## References

1. Javier Yetano Laguna and Vicent Alberola Cuñat. *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias*. Ministerio de Sanidad y Consumo, 2003.
2. Daniel Luna, Gastón Lopez, Carlos Otero, Alejandro Mauro, Claudio Torres Casanelli, and Fernán González Bernaldo de Quirós. Implementation of interinstitutional and transnational remote terminology services. In *AMIA Annual Symposium Proceedings*, volume 2010, page 482. American Medical Informatics Association, 2010.
3. Christopher Manning and Dan Klein. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5*, NAACL-Tutorials '03, pages 8–8, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
4. Hernán Navas, Alejandro Lopez Osornio, Analía Baum, Adrian Gomez, Daniel Luna, Fernan Gonzalez Bernaldo de Quiros, et al. Creation and evaluation of a terminology server for the interactive coding of discharge summaries. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 650. IOS Press, 2007.
5. Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
6. J. Pustejovsky, J. Castaño, B. Cochran, M. Kotecki, and M. Morrell. Automatic extraction of acronym-meaning pairs from medline databases. In *Proceedings of Medinfo, London*, 2001.
7. Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.