Biomedical Abbreviation Recognition and Resolution by PROSA-MED

Soto Montalvo[1], Maite Oronoz[2], Horacio Rodríguez[3], Raquel Martínez[4]

[1] URJC, soto.montalvo@urjc.es
[2] UPV/EHU, maite.oronoz@ehu.eus
[3] UPC, horacio@cs.upc.edu
[4] NLP&IR Group, UNED, raquel@lsi.uned.es

Abstract. The amount of abbreviations used in biomedical literature increases constantly. Despite the existence of acronym dictionaries, it is not viable to keep them updated with new creations. Thus, in the processing of biomedical texts, discovering and disambiguating acronyms and their expanded forms are essential aspects and this is the objective proposed by BARR task at IberEval 2017 Workshop. This paper presents our participation in this task. We propose five systems that deal with the problem in different ways. Three of the systems are atomic approaches, while two of them are combinations of the atomic systems. One of the systems clearly outperforms the others, both in the detection of entities (F-score of 0.749 in the test set) as well as identifying relations between short-long forms (F-score of 0.697 in the test set).

Keywords: abbreviation recognition, abbreviation disambiguation, patterns, dictionaries

1 Introduction

The volume of biomedical texts is greater and greater, and at the same time, the number of biomedical abbreviations is growing rapidly, being the ambiguity of biomedical abbreviations a challenge. Particularly, handling abbreviations without nearby definitions is a critical issue [3].

The acronyms have a high reference value, in the sense that they most of the time act as reference anchors of textual context [6]. Because of this and the common problem of recognition of abbreviations, acronyms and symbols, and their disambiguation (the same short form can have several different long forms), the Biomedical Abbreviation Recognition and Resolution (BARR) track is proposed [2].

Usually, existing work on acronym recognition in medical domain is proposed for English biomedical documents, being difficult to adapt these proposals to other languages. The BARR track has the aim to promote the development and evaluation of biomedical abbreviation identification systems in Spanish biomedical documents.

In this paper we present the approaches proposed by our team, in particular five different proposals.

The remainder of this paper is organized as follows. Section 2 presents the proposed systems. Section 3 summarizes the results and discuss about them. Finally, conclusions are presented in Section 4.

2 Proposed Systems

We propose different approaches to identify entities, both short and long forms, in the texts and also the relations between them. Our team is composed by members of different universities, in a way that we propose a method by each university (URJC and UNED propose one method together) and two additional methods which combine the other three proposal in some way. Following we describe all the methods.

2.1 EHU atomic approach

This system tries to take advantage of an already developed linguistic analyser, called FreelingMed [7]. This analyzer has been adapted to provide all the possible expansions for the abbreviations and acronyms that are already stored in its dictionaries. FreelingMed tokenizes the text, assigns the offsets to each token and identifies the medical terms appearing in SNOMED CT [11] as multiword terms. The output of the analyzer is usually given in XML but we have changed it to a format that is easier to manage (see Figure 1).

Fig. 1. The output of the FreelingMed analyzer.

Introducción	introducción	0:12
La	el	14:16
coexistencia	coexistencia	17:29
de	desviación_estándar # disfunción_erectil	30:32
esclerosis múltiple	esclerosis_múltiple	33:52
EM	electromiograma # eritema_multiforme # esclerosis_múltiple # estancia_media # estenosis_mitral	54:56

Figure 1 shows that for the acronym *"EM"* five possible expansions or medical terms are stored in the dictionaries of FreelingMed: *"electromiograma"*, *"eritema multiforme"*, *"esclerosis múltiple"*, *"estancia media"* and *"estenosis mitral"*. The EHU approach looks for all these expansions around the *"EM"* acronym, and if any of them is found, the relation is written as result. The whole medical term is searched near the short form as a unique unit but not the subelements of it (*"esclerosis"* or *"múltiple"*).

For the detection of entities, three main approaches are considered: i) an heuristic that marks word-forms that follow certain pattern usually appearing in abbreviations and acronyms; ii) elements that come marked as abbreviation or acronym from the dictionaries of FreelingMed; and, iii) elements that come from Freeling (in the basis of FreelingMed) marked as abbreviations referring to units of weight (e.g. *mg*), length (e.g. *cm*), time (e.g. *min*) etc.

2.2 UPC atomic approach

The second atomic system is based on the combination of three acronym / expansion pair extractors covering roughly the three most frequent cases of acronym / expansion mentions:

- *Similarity-based*. This approach tries to detect in a document (within the title and the abstract) mentions of single words or multiwords likely to be an acronym (short form) and an expansion (long form) so that the two forms are likely able to be mapped using a set of 13 hand crafted mapping rules. These mapping rules are applied in decreasing order of confidence. For recognizing the short forms we have used the set of regular expressions proposed in [5] constrained for satisfying the strict form of the word shapes proposed in [1][5]. For long form candidates we have collected all the ngrams up to 5 words, constrained for satisfying loose word shapes, and discarding the candidates starting or ending by a stopword. The set of allowed word shapes has been built from the annotations in the training set. The most frequent and most accurate rule can be paraphrased as following: "The length of the acronym in chars has to be equal to the number of expansion tokens. Each character of the acronym should correspond to the first letter of the corresponding token in the expansion".
- *Gazetteer-based*. We have used a big terminology of the medical domain obtained from several sources (containing 103,169 terms). The terminology, which covers six languages was compiled following an iterative approach in a way that at each iteration available resources for one language were included and then mapped, when possible, to other languages using *dbpedia* links ("sameAs" and "label"). The main source of resources includes for English *Bioportal*[6] and *DrugBank*[7], for Spanish *CIE10*[8] and *CIMA*[9], and for French *pyMedTermino*[10]. The terminology includes both short forms and long forms and we have obtained possible pairs using the *Similarity-based* approach described above. 14,360 pairs were obtained in this way. An example of such

[5] A word shape is a simple pattern aiming to represent the character level form of a word (case, letter, number, punctuation mark, space), e.g. the strict form shape of 'DM2' is 'AA0' while the loose shape is 'A0'.
[6] https://bioportal.bioontology.org/
[7] https://www.drugbank.ca/
[8] CIE10.org
[9] https://www.aemps.gob.es/cima/
[10] https://bitbucket.org/jibalamy/pymedtermino

patterns (represented as a regular expression) is u' (LPC) .0,15 (linfoma primario cerebral) '.
- *Distance-based.* We have collected a set of patterns acronym / expansion occurring closely and frequently in the training set. The most frequent pattern is represented by the regular expression ([A-Za-z][^]+ [A-Za-z][^]+ [A-Za-z][^]+) ($[A-Z]3,3$), covering, for instance, "enfermedad renal crónica (ERC)". This pattern occurs 73 times in the training set.

2.3 UNED atomic approach

This system combines a pattern-based approach with a dictionary-based approach, and consists on two steps: abbreviations detection and definition matching for them.

In the first step, we detect terms in capital letters or combination of capital letters with lowercased letters, numbers and other characters. We use parenthetical constructions as indicator of a possible abbreviation [8]. Once the abbreviation (short form) is located and validated, the second step searches for its definition (long form) on the left side of the open parenthesis using the algorithm proposed by Schwartz and Hearst [10]. We select each word, one by one and combining them in each iteration, until a combination of them match with the short form. We have extended the algorithm of Schwartz and Hearst in order to allow the words of the long form do not appear necessarily in the same order that the characters of the short form. The number of words we combine searching the long form do not exceed the double of the characters of the short form.

In addition, some special cases for the approach based on patterns are considered. For instance, the following text has two relation pairs for the same acronym and two different definitions, one per language: *"amino-terminal propeptide of procollagen type 1 (P1NP, propéptido aminoterminal del procolágeno 1)"*.

In case of the pattern-based approach does not find a valid definition for the abbreviation, we use a dictionary where each entry is an abbreviation and its possible long forms. In the same order that long forms appear in the dictionary, we search each one in the same sentence where the abbreviation is, and the first one that matches is selected as the long form of the pair of the relation. The dictionary used has 7,916 entries.

2.4 Output Combination

We have implemented three simple combination mechanisms named as *and*, *or*, and *vot*, that are applied over the results of the atomic systems. *and* accepts an annotation only in the case all three atomic systems propose it. *or* accepts all the annotations of the atomic systems just checking that no contradictions (partial overlapping) occur in the mentions. *vot* implements a democratic votation schema, i.e. an annotation is accepted in the case at least two of the atomic systems have proposed it.

For our final submission only *and* and *or* combinations were submitted.

3 Results and Discussion

In this section we present the results obtained identifying entities and relations abbreviation-definition.

The evaluation metric used for evaluating the participating systems of the BARR track has been the F-score micro measure. The organization has provided an adaptation of the Markyt platform for the evaluation [9]. This platform allows to visualize and compare generated predictions against the Gold Standard annotations.

The organization has provided training, test and background collections [4]. Table 1 shows the results of the six systems over the training 1 data set. The first column shows the system, and the columns 2-4 and 5-7 show the values of precision, recall and F-score respectively, for the identification of Entities and Relations. On the other hand, the three first rows show the results for the atomic systems, and the last three rows the results for the systems that combine the previous ones. In both cases the systems are ordered by F-score value.

Table 1. Preliminary results of the proposed systems over training 1 set.

	Entities			Relations		
System	**P**	**R**	**F**	**P**	**R**	**F**
UNED	0.86	0.64	0.74	0.79	0.58	0.67
UPC	0.30	0.39	0.34	0.34	0.30	0.32
EHU	0.29	0.37	0.33	0.90	0.10	0.18
Comb OR	0.32	0.47	0.38	0.39	0.38	0.39
Comb AND	1.00	0.06	0.12	1.00	0.06	0.12
Comb VOT	0.48	0.08	0.14	1.00	0.07	0.13

The UNED system obtains high precision values, specially identifying entities. This confirm that an approach based on patterns is suitable for this problem. The recall values are a bit lower due to this system does not detect nested entities. Moreover, it is probably the patterns did not detect all special cases that could appear in texts. The system considers some special cases which implies variations in the patterns, but it is possible that exist more special cases not considered.

The F-score is lower identifying relations because not for all entities detected the system finds a valid long form. The system searches long forms in a maximum number of words on the left of an acronym and it can be out of this window.

The main objective of the EHU system has been to reuse a linguistic analyzer that was already developed. This approximation is limited as it only can detect abbreviations already gathered in the dictionaries of FreelingMed and analyzed as an unique element in its long form. Table 1 shows in its relation column that a recall of 0.1 is obtained but with a precision of 0.9. Those results, in our opinion, are clearly related to the type of approach.

The results of UPC system were bad. There are several explanations for this:

- The *Gazetteer-based* component had a very small contribution to the global system.
- The acronym detector resulted in the training phase on many failures (about 50 false negatives and more than 200 false positive). Specially in the case of one character abbreviations the results were bad.
- Also the detection of long form candidates resulted in many false positives, specially single word terms and multiterms starting or ending with a non valid POS.
- Finally, some of the mapping rules, specially those involving a single word long form, presented a low accuracy.

Table 2 presents the results over the test set. For the test set only the *and* and *or* combinations were done between the UNED and UPC systems output.

Table 2. Results of the runs over test set.

	entities			relations		
System	**P**	**R**	**F**	**P**	**R**	**F**
UNED	0.87	0.66	**0.75**	0.81	0.60	**0.70**
UPC	0.65	0.21	0.32	0.31	0.10	0.15
EHU	0.25	0.10	0.15	0.61	0.02	0.03
Comb OR	0.77	0.37	0.50	0.57	0.27	0.36
Comb AND	0.99	0.10	0.18	0.98	0.09	0.16

Due to the big volume of data in the test set, only the 22.5 % of the corpus was analyzed in time in the EHU approach and there were some computer memory problems in the UPC approach. FreelingMed is quite slow (39 sec. to analyze an abstract of 178 tokens) due to the volume of the dictionaries it uses: a token dictionary of 578,539 entries and a multiword dictionary of 474,800 entries. To detect words usually used with a non-medical meaning, for instance *"bar"*, with a medical meaning (e.g. *"bacilo acidorresistente"*), a second analysis phase is applied with a dictionary of around 930,000 entries. In addition a mapping between SNOMED-CT and the Unified Medical Language (UMLS[11])(1,007,705 entries) is applied. Not all these resources are needed for the BARR task, but they are already included in FreelingMed. The time problem with FreelingMed and the memory problems in the UPC approach have a direct relationship with the results in the recall column shown in Table 2.

As can be seen on both tables (Tables 1 and 2), the UNED approach outperforms by large extent every other one for both tasks and all measures. Taking into account this high difference, combinations produce no improvement. It is

[11] https://www.nlm.nih.gov/research/umls/

worth noting, however, that *and* combination reach the best precision for both tasks, at a cost of a extremely low recall. The NESTED type has not been treated in the entities identification task.

4 Conclusions

This paper has described our participation in the BARR task at IBEREVAL 2017 workshop, which goal is to find acronyms and acronyms-long form relations. We have proposed five different approaches, three atomic systems and two more systems, which combine on different ways the atomic proposals.

The UNED system clearly stands out among the presented systems. Being so clear the difference in results with the two other atomic approaches, the combinations are not able of improving the UNED system results.

Dictionary-based approaches are language dependent while the ones based on the use of regular expressions or patterns show to be more flexible.

There is, obviously, room for improvements. We plan to focus on performing combination not only as a final process but using partial results from the other atomic sources.

5 Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation (PROSA-MED Project: TIN2016-77820-C3, TADEEP Project: TIN2015-70214-P).

References

1. Dai, HJ. & Lai, PT. & Chang, YC. & Tzong-Han Tsai, R.: Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. Journal of Cheminformatics 2015 7(S-1) (2015)
2. Intxaurrondo, A. & Pérez-Pérez, M. & Pérez-Rodríguez, G. & Lopez-Martin, J.A. & Santamaría, J. & de la Peña, S. & Villegas, M. & Akhondi, S.A. & Valencia, A. & Lourenço, A. & Krallinger, M.: The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts. SEPLN 2017, (2017)
3. Kim, S. & Yoon, J.: Link-topic model for biomedical abbreviation disambiguation. Journal of Biomedical Informatics,53:367-380. (2015)
4. Krallinger, M. & Intxaurrondo, A. & Lopez-Martin, J.A. & de la Peña, S. & Pérez-Pérez, M. & Pérez-Rodríguez, G. & Santamara, J. & Villegas, M. & Akhondi, S.A. & Lourenço, S. & Valencia, V.: Resources for the extraction of abbreviations and terms in Spanish from medical abstracts: the BARR corpus, lexical resources and document collection. SEPLN 2017, (2017)
5. Larkey, L.S & Ogilvie, P. & Price, M.A. & Tamilio, B.: Acrophile: an automated acronym extractor and server. Proceedings of the fifth ACM conference on Digital libraries (ACM DL), pp. 205-214. (2000)

6. Maud, E. & della Rocca, L. & Steinberger, R. & Tanev, H.: Acronym recognition and processing in 22 languages. Proceedings of the 9th Conference Recent Advances in Natural Language Processing (RANLP), pp. 237244. (2013)
7. Oronoz, M. & Casillas, A. & Gojenola, K. & Pérez, A.: Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013 Havana, Cuba, November 20-23. (2013)
8. Park, Y. & Byrd, R.J.: Hybrid Text Mining for Finding Abbreviations and Their Definitions. Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, pp. 126-133. (2001)
9. Pérez, M., Pérez, G., Rabal, O., Vazquez, M., Oyarzabal, J., Fernández, 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)
10. Schwartz, A.S & Hearst, M.A.: A simple algorithm for identifying abbreviations definitions in biomedical text. Pacific Symposium on Biocomputing, pp. 451-462. (2003)
11. SNOMED-CT, Systematized Nomenclature of Medicine-Clinical Terms. International Health Terminology Standards Development Organisation (IHTSDO). 2016. Accessed 2014-04-09.