MAMTRA-MED at Biomedical Abbreviation Recognition and Resolution - IberEval 2018

Soto Montalvo[1], Raquel Martínez[2], Mario Almagro[2], Susana Lorenzo[3]

[1] URJC, soto.montalvo@urjc.es
[2] NLP&IR Group, UNED, raquel@lsi.uned.es
[3] NLP&IR Group, UNED, malmagro@lsi.uned.es,
[4] HUFA, slorenzo@fhalcorcon.es

Abstract. Biomedical literature, Electronic Health Records and clinical notes frequently contain abbreviations and acronyms, most of them are highly ambiguous. Correct interpretation of abbreviations and acronyms often represents a challenge, and this is the objective of the BARR2 track at IberEval 2018 Workshop. This paper presents our participation in the track. We propose five systems that deal with the detection of explicit occurrences of abbreviations-definitions pairs in different ways (sub-track1), and three systems to resolve abbreviations and acronyms regardless of whether its definition is mentioned within the actual document (sub-track2). The proposals for detecting relations between abbreviations and their definitions have obtained good results, while the proposals for abbreviation resolution, the more complex of the two tracks, has great room for improvement.

Keywords: abbreviation recognition, abbreviation resolution, abbreviation disambiguation, patterns, dictionaries

1 Introduction

The number of biomedical texts is growing continuously, with frequent use of abbreviations and acronyms. These short forms often contain important clinical information (eg, names of diseases, drugs, or procedures) that must be recognizable and accurate in health records [11]. Understanding these clinical short forms is still a challenging task for current clinical natural language processing (NLP) systems [10].

Many abbreviations are ambiguous because they have more than one possible expansion. For example, expansions for "om" include "oído medio", "orificio mitral", "obesidad mórbida", and "osteomielitis", among others. The accurate recognition of abbreviation sense is largely significant to understand and analyze biomedical data [4]. Particularly, handling abbreviations without nearby definitions is a critical issue [5].

Most proposals for recognizing and disambiguating short forms are applied over English biomedical texts. First edition of the BARR track [3] appeared with the aim of promoting the development and evaluation of biomedical abbreviation

identification systems in Spanish biomedical documents. The Second Biomedical Abbreviation Recognition and Resolution track (BARR2) [2] continues with the same aim and extends annotations of short forms to clinical text and clinical case studies written in Spanish. Two sub-tracks have been proposed, one for detecting only explicit occurrences of abbreviations-definitions pairs, and the other to associate definitions with short forms even if they don't appear in documents. In this paper we present different approaches for each sub-track, tacking the advantage of our participation in the previous edition of BARR track [9].

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

2 Proposed Systems

We propose different approaches to identify explicit occurrences of short form-long form pairs and also to disambiguate short forms. Next subsections present these approaches. Previously, documents are split into sentences, and for each of them the short forms are detected, both to find their long forms and also to disambiguate them.

2.1 Abbreviations-definitions pairs detection systems

We propose five systems to detect the abbreviations-definitions pairs in the clinical texts: a pattern-based approach, a dictionary-based approach and the other three systems are combinations of the first ones. In general, these systems consists of two steps: abbreviations detection and definition matching for them. In the first step, we detect terms in capital letters or combinations of capital letters with lowercased letters, numbers and other characters.

In our pattern-based approach ($RUN1_R$) we use parenthetical constructions as indicator of a possible abbreviation or acronym [6]. Once the 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 [7]. We select each word, one by one and combine them in each iteration, until a joining matches with the short form. We have extended the algorithm of Schwartz and Hearst in order to allow the words of the long form not necessarily to appear in the same order that the characters of the short form. The number of words we combine by searching the long form do not exceed the double of the characters of the short form.

In addition, some special cases in the approach based on patterns are considered. For instance, sometimes the definition appears in parenthesis instead of the acronym itself (*DBX (matriz sea desmineralizada)*), or the acronym do not have any character in uppercase (*lpm (latidos por minuto)*).

In our dictionary-based approach ($RUN2_R$), texts are tokenized first, and then we look for a short form in our dictionary for considered token, where each

entry is an abbreviation and its possible long forms. If the dictionary contains the short form then system searches one of the long forms associated with it in the text; finally, the first one that matches is selected as the long form of the pair of the relation.

Our $RUN3_R$ and $RUN4_R$ systems combine the two previous approaches. On one hand, $RUN3_R$ system prioritizes the relations found by the dictionary-based approach, while $RUN4_R$ system prioritizes the relations found by the pattern-based approach. Finally, the $RUN5_R$ system combines the two first approaches, but in a different way. In case of the pattern-based approach, if it does not find a valid definition for the short form, so we use the dictionary for searching a valid long form.

2.2 Abbreviation disambiguation systems

We propose three systems for the abbreviation disambiguation, all of them based on the use of two repositories or dictionaries: one is the dictionary used in the previous sub-track and the other consist of a short dictionary with abbreviations of measurement units of the medical domain.

The system named $RUN1_D$ first obtains a list of abbreviations-definitions pairs with the proposed $RUN4_R$ system, because this proposal obtained good results with the training and development data. Next, texts are tokenized and the abbreviations contained in the list of relations are directly disambiguated with the long form of their relation. In other case, it searches the token in the measurements units dictionary: if find it then disambiguate the abbreviation with the definition associated in the dictionary; otherwise, it searches the token in the other dictionary. This last one contains one or more definitions for each abbreviation depending on the ambiguity, and in some cases, also contains text related with definitions (next section explains the content of the dictionary and how it was generated). If an abbreviation is found in the dictionary and only contains a long form, this will be the assigned definition; otherwise,if there are several definitions, the approach will return the definition more frequent, that is, the definition with more texts associated. The text associated with a definition belongs to abstracts of biomedical articles where the abbreviation has appeared.

The $RUN2_D$ system only differs from the previous approach when an abbreviation is ambiguous. In this case, instead of disambiguating it with the most frequent definition, it chooses the definition whose tokens appear most frequently in the clinical case text in which the ambiguous abbreviation appears. Finally, our third proposal ($RUN3_D$) is the same as $RUN2_D$ system, but also including the context of a definition when checks the frequency of their tokens in the original text of the ambiguous abbreviation.

3 Experiments

In this section we describe the dictionaries used in our systems as well as the results obtained from them.

3.1 Dictionaries

We have created two dictionaries, one with abbreviations that represents units of measurements, and other with abbreviations that appear usually in biomedical texts. The first dictionary is small, containing only 89 abbreviations, while the second one contains 7169 abbreviations, 2531 of them ambiguous.

The measurements units dictionary was created looking up in web pages related with medical organizations, care centers or regional recommendations about the medical terminology. On the other hand, our bigger dictionary of abbreviations was created considering different sources: SNOMED [8], a Spanish dictionary of abbreviations [12], Spanish abbreviations from Wiki LaEnfermería[5], and abbreviations about rare diseases compiled from Orphanet[6]. In addition, we have processed some of the data provided in the BARR track [3], in particular the training1, training2 and sample data sets, in order to extract more abbreviations and their definitions, and also to extract their contexts. These contexts are the biomedical texts in which the abbreviations appear, and those we use in some of our systems for abbreviations resolution.

3.2 Results and Discussion

In this section we present the results obtained by identifying abbreviations-definitions pairs and also those from the abbreviation disambiguation.

The evaluation metric used for the BARR2 track consists of micro-average F-measure of explicit occurrences of abbreviations-definitions pairs. F-measure stands for the harmonic mean between precision and recall. Individual scores for each abbreviation-definition pair are computed in different ways for sub-track 1 and 2. More details about the computation of F-measure in each case can be found in the overview of the BARR2 track [2].

The organization has provided training, development, and background and test collections [1], although the annotations for the two sub-tracks are only provided for the two formers. Table 1 shows the results of the five systems over the training and development data sets for the abbreviations-definitions pairs detection, and Table 2 shows the results of the three systems over the training and development data sets for the abbreviation resolution sub-track. In both tables, the first column shows the system, while columns 2-4 and 5-7 show the values of precision, recall and F-score respectively, both for the training and development sets.

The evaluation script provided by the organization for evaluating abbreviation resolution sub-track runs three different evaluations: ultra-strict, strict and flexible evaluation. F-measure values showed in Table 2 correspond to a flexible evaluation.

In general, all the proposed systems for the detection of abbreviations-definitions pairs achieve high precision values. This confirm that approaches based on patterns or dictionaries are suitable for this problem, specially if they are combined.

[5] http://www.laenfermeria.es/docuwiki/doku.php?id=siglas_medicas
[6] https://www.orpha.net/consor/cgi-bin/index.php

Table 1. Preliminary results of the proposed systems for abbreviations-definitions pairs detection over training and development sets.

	Training set			Develop. set		
System	**P**	**R**	**F**	**P**	**R**	**F**
$RUN1_R$	0.85	0.76	0.80	0.87	0.74	0.80
$RUN2_R$	0.92	0.44	0.60	0.92	0.48	0.63
$RUN3_R$	0.85	0.77	0.81	0.85	0.77	0.81
$RUN4_R$	0.85	0.77	0.81	0.87	0.78	0.82
$RUN5_R$	0.85	0.77	0.81	0.87	0.74	0.80

Table 2. Preliminary results, by means of a flexible evaluation, of the proposed systems for abbreviation resolution over training and development sets.

	Training set			Develop. set		
System	**P**	**R**	**F**	**P**	**R**	**F**
$RUN1_D$	0.43	0.25	0.32	0.42	0.27	0.33
$RUN2_D$	0.40	0.24	0.30	0.39	0.25	0.30
$RUN3_D$	0.44	0.26	0.33	0.43	0.28	0.34

The $RUN2_R$ system, it means, the approach based on the use of a dictionary obtains the highest precision values because in this case the probability of selecting a wrong definition is lower. If the definition is in the dictionary, it will always be selected, but in the pattern-based approach is more probable to select wrong definitions with more o less words than the correct one. On the other hand, the recall values are a bit lower because none of the systems detect nested entities, so that our systems only detect short-long pairs, but not short-nested nor nested-long pairs. Moreover, it is probably that the patterns did not detect all special cases that could appear in texts. In particular, the worse recall values are obtained by the $RUN2_R$ system, because in this case only the abbreviations found in the dictionary are considered. Therefore there are more different abbreviations on texts than those contained in the dictionary.

Our proposals for disambiguating abbreviations have not obtain good results. It seems that approaches based only on dictionaries is not enough, because if the abbreviation detected is not in the dictionary it cannot be disambiguated. Although this is not the unique reason of the poor results, there are errors in the detection of abbreviations on texts, due to the high variety of them, and specially to those that represent measures. For example, the tokens *10mg/dl* or *3,24mg/dl* contain two abbreviations, or the token *23ml/min/1,73m2* contains three. Therefore, it is necessary an extra processing different in each case.

Table 3 presents the results over the test set, which are similar or a bit worse than results obtained over training and development sets.

Table 3. Results of the runs over test set, for abbreviations-definitions pairs detection and for abbreviation resolution tasks.

System	P	R	F
$RUN1_R$	0.85	0.70	0.76
$RUN2_R$	0.91	0.47	0.62
$RUN3_R$	0.85	0.73	0.79
$RUN4_R$	0.84	0.73	0.78
$RUN5_R$	0.85	0.70	0.77
$RUN1_D$	0.39	0.22	0.28
$RUN2_D$	0.37	0.21	0.27
$RUN3_D$	0.42	0.24	0.31

4 Conclusions

This paper has described our participation in the BARR2 task at IBEREVAL 2018 workshop, whose goal is to find abbreviations-definitions relations and resolution of abbreviations in Spanish biomedical texts. We have proposed five different approaches for the first sub-track, two atomic systems and three more systems, which combine on different ways the atomic proposals, and we have proposed three approaches for the second sub-track.

Our proposals for detecting abbreviations-definitions pairs obtain good results, but we have room for improvement. There are some special cases on texts that our patterns have not identified, as well as our systems have not detected nested relations, which is one of the requirements for the predictions. These aspects could be improved in the future. On the other hand, we would like to improve our systems for abbreviations resolution because our results are poor. Our approach based on patterns and dictionaries is not enough to resolve abbreviations, therefore we could apply other techniques in order to complement the approach with dictionaries.

5 Acknowledgments

This work has been funded by the Spanish Ministry of Science and Innovation (MAMTRA-MED Project, TIN2016-77820-C3-2-R and MED-RECORD Project, TIN2013-46616-C2-2-R).

References

1. Intxaurrondo, A. & de la Torre, J.C. & Rodríguez Betanco, H. & Marimon, M. & López-Martín, J.A. & González-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. SEPLN 2018

2. Intxaurrondo, A. & Marimon, M. & González-Agirre, A. & López-Martín, J.A. & Rodríguez 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. SEPLN 2018.
3. Intxaurrondo, A. & Pérez-Pérez, M. & Pérez-Rodríguez, G. & López-Martín, 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)
4. Kai, R & Shi-Wen, W.: Applying Convolutional Neural Network Model and Auto-expanded Corpus to Biomedical Abbreviation Disambiguation. Journal of Engineering Science and Technology Review, 9(6): 178-184 (2016)
5. Kim, S. & Yoon, J.: Link-topic model for biomedical abbreviation disambiguation. Journal of Biomedical Informatics,53:367-380. (2015)
6. 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)
7. Schwartz, A.S & Hearst, M.A.: A simple algorithm for identifying abbreviations definitions in biomedical text. Pacific Symposium on Biocomputing, pp. 451-462. (2003)
8. SNOMED-CT, Systematized Nomenclature of Medicine-Clinical Terms. International Health Terminology Standards Development Organisation (IHTSDO). 2016. Accessed 2014-04-09.
9. Montalvo, S., Oronoz, M., Rodrguez, H., Martnez, R.: Biomedical Abbreviation Recognition and Resolution by PROSA-MED. Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017. CEUR Workshop Proceedings 1881, CEUR-WS.org 2017: 247-254.
10. Wu, Y. & Denny, J.C. & Rosenbloom, S.T. & Miller, R.A. & Giuse, D.A. & Wang, L. & Blanquicett, C. & Soysal, E. & Xu, J. & Xu, H. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). Journal of the American Medical Informatics Association, 24(e1): e79e86(2017)
11. Xu, H. & Stetson, P.D. & Friedman, C.: A study of abbreviations in clinical notes. AMMIA Annual Symposium Proceedings, pp. 821-825. (2007)
12. Yetano Laguna, J., Alberola Cuat, V.: Diccionario de siglas médicas. SEDOM, Sociedad Española de Documentación Médica. (2012)