IBI-UPF at BARR-2017: learning to identify abbreviations in biomedical literature
System description

Francesco Ronzano and Laura I. Furlong

Integrative Biomedical Informatics Group, Research Programme
on Biomedical Informatics (GRIB)
Hospital del Mar Medical Research Institute (IMIM)
Universidad Pompeu Fabra
Barcelona, Spain
{francesco.ronzano,laura.furlong}@upf.edu

Abstract. This paper presents the participation of the IBI-UPF team to the Biomedical Abbreviation Recognition and Resolution (BARR) track organized in the context of the Evaluation of Human Language Technologies for Iberian Languages 2017 (IBEREVAL). The purpose of the track was to automatically identify abbreviation-definition pairs in the abstract of biomedical articles in Spanish. By releasing a sample corpus and two collections of training documents, the organizers provided a total of 1,150 abstracts of biomedical articles, the majority of them in Spanish, manually annotated with respect to the identifications of abbreviations and the corresponding definitions. We tackled the task by implementing an approach articulated in two sequential phases. In the first one, by relying on a set of shallow linguistic features extracted from the textual contents of biomedical abstracts, we trained two token classifiers to spot sequences of one or more tokens that respectively represent abbreviations or definitions. Then, a third classifier is trained to distinguish abbreviations that are candidate short forms of a definition expressed in the same abstract sentence from other types of abbreviations. In a second phase, relations between the abbreviations and definitions previously spotted are identified by means of a set of heuristics based on structural and linguistic traits of the text of each abstract. We evaluate the first phase of our approach by considering the set of Spanish biomedical abstracts manually annotated, provided by the organizers of the BARR track.

1 Introduction

Nowadays, automated approaches to mine biomedical texts are becoming key tools to enable researchers, as well as any other interested actor, to effectively access to and take advantage of the huge and rapidly growing amount of articles available on-line [6]. PubMed[1], the main search engine of life science and biomedical papers, currently includes more than 27 million articles and is growing at a rate of about 7% of new publications every [18].

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

Abbreviations, acronyms and symbols are extensively used in biomedical texts: their identification and correct interpretation are essential to automatically analyze this kind of documents. Several approaches have been proposed during the last decades to extract abbreviation-definition pairs in biomedical texts [9, 17]. Part of them are based on a mix of pattern-matching and heuristic rules sometimes complemented by corpus statistics [2, 7, 14, 20, 22, 23] while other ones propose hybrid systems that rely on supervised learning approaches that are properly trained on manually annotated corpora [3, 12, 13, 21]. During the last decade, in the biomedical domain, besides scientific papers, also clinical notes have focused several efforts towards the autormated extraction and interpretation of abbreviations [4, 11, 19].

The Biomedical Abbreviation Recognition and Resolution (BARR) [8] track has been organized in the context of the Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL 2017) in order to promote the investigation of new approaches to identify abbreviations together with their definitions in Spanish biomedical documents. In this paper we describe our participation (UPF-IBI team) to the BARR track. In particular, in Section 2 we provide more details on the BARR task by introducing some core aspects of the BARR corpus of biomedical abstracts manually annotated with respect to abbreviations. In Section 3 we describe the set of Natural Language Processing tools and resources we exploited to support the automated identification of abbreviation-definition pairs in biomedical abstracts. Section 4 explains our approach to face the BARR task. In Section 5 we provide some preliminary evaluation of our automated abbreviation identification system by considering the training set of manually annotated abstracts provided by BARR organizers. To conclude, in Section 6 we summarize the key points of our BARR participation outlining future venues of research to improve our approach.

2 BARR track: task and dataset

The information extraction task proposed to the participants of the BARR track consists in the identification of abbreviations (or Short Forms, *SF*s) that occur in sentences of Spanish biomedical abstracts and their association to the corresponding definitions, referred to as Long Forms (*LF*s). An example of ⟨*SF, LF*⟩ pair is ⟨*TAC, Tomografa Axial Computarizada*⟩. Besides proposing approaches to mine the broad variety of possible *SF*s that can be exploited to refer to a specific *LF*, BARR participants were also required to deal with the detection of nested ⟨*SF, LF*⟩ pairs: in these pairs two or more *SF*s share portions of the corresponding *LF*s or the *LF* associated to a *SF* is not constituted by a consecutive sequence of words. The expression *dolor oncológico (DO) y no oncológico (DNO)* includes two nested ⟨*SF, LF*⟩ pairs: ⟨*DO, dolor oncológico*⟩ y ⟨*DNO, dolor no oncológico*⟩.

In order to train automated approaches for the detection of ⟨*SF, LF*⟩ pairs (both simple and nested ones), BARR organizers released a sample corpus and two training corpora globally providing 1,150 manually annotated abstracts of biomedical articles: about 90% of these documents are Spanish texts. The evaluation of the abbreviation extraction approaches proposed in the context of the BARR task is performed by computing precision, recall and f1-score of each proposed approach with respect to a test

corpus that includes 600 Spanish biomedical abstracts: the extraction of entities (*SF*s and *LF*s) and their relations are considered as two separate tasks. More details concerning the corpus of biomedical papers released in the context of the BARR track together with the description of how these documents have been manually annotated can be found in [10].

3 Tools and resources

To identify *SF*s, *LF*s and their associations, we exploited a mix of machine learning and heuristic approaches, both based on the characterization of the textual contents of biomedical abstracts through a set of shallow linguistic and corpus-based features. We computed these features by processing Spanish abstracts by means of the IXA Pipes NLP tools [1]: we performed sentence splitting, tokenization, Part of Speech tagging and constituency parsing. To process Spanish documents, IXA Pipes rely on NLP models trained on the Spanish texts of the AnCora Corpus[2]. Besides linguistic analyses, we determined the frequency of usage of abstracts' words by relying on a word-frequency dictionary built from a 2016 dump of the Spanish Wikipedia. We exploited the GATE Framework [5] to integrate the text mining tools just mentioned into a single pipeline.

4 Method

Our abbreviation identification approach is composed of two sequential steps: the entity spotting and the relation extraction phase. The first phase relies on machine learning approaches to identify and characterize both *SF*s and *LF*s. The second phase exploits a set of heuristics in order to refine the entities previously identified and extract relations between *SF*s and *LF*s. We considered among the heuristics implemented in the second phase, a set of rules properly built to automatically characterize simple cases of nested $\langle$*SF, LF*$\rangle$ pairs. In this Section we provide a detailed description of the two phases of our abbreviation identification approach.

4.1 Phase 1: entity spotting

The first phase of our approach aims at: (i) extracting abbreviations and *LF*s; (ii) selecting, among the spotted abbreviations, the ones that are *SF*s and thus occur in the same sentence of the corresponding *LF*.

All these information extraction tasks have been performed by training distinct token-based classifiers. In these classifiers each token is characterized by means of the following types of features that we exploited to model both the token under consideration and the ones included in a context window of size $[-2, 2]$:

- Part of Speech;
- number of characters, including punctuations;
- percentage of uppercase, numeric and punctuation characters;

[2] http://clic.ub.edu/corpus/ancora

- if the first / last char is uppercase;
- if the last char is a punctuation;
- number of repetitions of the token in the abstract;
- match of the token with one of the entries of the Dictionary of Medical Abbreviations SEDOM[3];
- frequency of the token in the Spanish Wikipedia.

Each one of the types of features listed before generates five feature values for each token: one describing the token under analysis and four characterizing respectively the two previous tokens and two following tokens in the same sentence. We plan to explore in our future work the influence of different window sizes on the performance of our token-based classifiers, by considering also windows that are symmetric and not-symmetric with respect to the token to classify. We computed token features scoped to each sentence, thus setting as missing the feature values of the context tokens that cannot be determined since they are outside sentence boundaries. We selected our set of features in order to describe traits of tokens and their context that we considered relevant to the identification and characterization of abbreviations and *LF*s. For instance the presence of high percentages of uppercase letters is proper of many abbreviations.

By relying on the previous set of features we build three Random Forest classifiers respectively trained to determine:

- **Abbreviation Token Classifier**: if a token represents or not an abbreviations;
- **Long Form Token Classifier**: if a token is at the Beginning, Inside or Outside a *LF*;
- **Abbreviation Type Classifier**: if a token classified as an abbreviation by the Abbreviation Token Classifier is a *SF* or represents another kind of abbreviation (e.g. an abbreviation for which the Long Form is not provided in the same sentence).

In our approach presented to the BARR track, after selecting the best subset of features with respect to the task to perform, we trained each classifier over the whole set of tokens of the manually annotated Spanish biomedical abstracts provided by the BARR track organizers. Section 5 includes an initial evaluation of the performance of our classifiers over the BARR manually annotated Spanish abstracts.

4.2 Phase 2: relation extraction

Once identified *SF*s and *LF*s, in this phase we mainly implemented the following set of heuristics to determine if a *SF* includes the related *LF* in the scope of the sentence where it occurs:

- **Long Form sanitizing heuristics**:
  (A.1) delete all *LF*s that have all tokens with a length shorter than three characters or that does not include a noun token;
  (A.2) remove the initial token from the text span of the *LF*s that start with an article;

[3] http://www.sedom.es/diccionario/

– ***SF - LF* relations identification heuristics**:

  (B.1) collect for each *SF* all the candidate *LF*s, including the *LF*s identified by the classifier and the noun phrases occurring in the same sentence, not overlapping the *SF*, distant from the *SF* at most three characters and spanning a number of characters bigger than the number of characters of the *SF*. If the *SF* is between parenthesis, we consider only the preceding candidate *LF*s;

  (B.2.1) if there is only one candidate *LF*: if the candidate *LF* has been identified by the Long Form Classifier, create a *SF - LF* relation. Otherwise, if it is a noun phrase apply the *SF-LF scoring function* (described below) and create a *SF - LF* relation if the score is greater than 0.

  (B.2.2) if there is more than one candidate *LF*: score each candidate *LF* by means of the *SF-LF scoring function* and chose the one with highest score, greater than 0. If there is more than one candidate *LF* characterized by the highest score give precedence to the one that has been identified by the Long Form Classifier, if any, otherwise choose one of the candidate *LF*s randomly.

As mentioned in the previous procedure, we defined a *SF-LF scoring function* that, given a pair of *SF* and candidate *LF*, returns a double value that is equal to 0 if the *LF* is not recognized as related to the *SF*. Otherwise such function returns a number greater than 0: the greater is this value, the higher we estimate that the candidate *LF* represents a definition of the *SF*. A value equal to 1 spots a perfect match between the *SF* and candidate *LF*. The return values of the *SF-LF scoring function* have been defined by relying on the precision estimates of the *SF* / *LF* matching strategies defined by [16].

We extended the *SF - LF* relation extraction procedure just described by means of a set of refinement steps so as to properly deal with special cases including:

– groups of *SF*s like *fibrosis intersticial y atrofia tubular [FI y AT]*.
– if no *LF* has been found, starting from the considered *SF* we try to build the *LF* by matching word-initials backwards;
– if no *LF* has been found, if the *SF* matches some of the abbreviations of the Dictionary of Medical Abbreviations SEDOM, we search for the corresponding *LF* retrieved from the same Dictionary in the set of candidate *LF*s previously described. This approach covers borderline cases like ⟨*CO2, Dixido de carbono*⟩ in which it would have been impossible to determine the *SF - LF* relation.

We also defined a basic set of heuristics to spot cases of nested relations between *SF* and *LF*s. We identify the eventual presence of nested relations if, after a candidate *LF* two or more *SF*s are present before the end of the sentence or the occurrence of the following candidate *LF*. If this situation occurs we exploit a set of rules based on string matching and POS tags so as to identify the NESTED entities and the *SF - NESTED* relations. In partiuclar, for each *SF* marked as nested candidate, we search backwards for non consecutive words matching the initials of the same *SF* and including at least one noun token.

5 Evaluations and runs

We evaluated the performance of the three Random Forest classifiers described in Section 4.1 by means of a 10-fold-cross-validation over the 237,603 tokens of manually annotated BARR abstracts (Table 1).

Classifier	*Precision*	*Recall*	*F1-score*
Abbreviation Token	0.937	0.918	0.927
Long Form Token	0.623	0.345	0.444
Abbreviation Type	0.838	0.828	0.833

Table 1. Evaluation of entity spotting phase classifiers over manually annotated BARR abstracts (micro-average): (i) Abbreviation Token Classifier: weighted F1-score of classification of tokens as abbreviation or not, (ii) Long Form Token Classifier: weighted F1-score of Beginning and Inside tokens, (iii) Abbreviation Type Classifier: weighted F1-score of classification of abbrev. in: DERIVED, GLOBAL, NONE, MULTIPLE, SHORT

From Table 1 we can notice that the identification and characterization of abbreviations obtain satisfactory performance. As far as concern the identification of *LF*s, the Random Forest classifier obtains a low F1-score. This drawback of the first processing phase of our system (Section 4.1), probably related to the need to define better token level features for *LF* identification, is mitigated by the second phase (Section 4.2) in which the *LF*s spotted by the Long Form Token Classifier are sanitized and properly complemented by the *LF* candidates retrieved by considering nominal phrases.

We submitted to the BARR track three runs to the entity extraction task and three runs to the relation extraction task (referred to as run v1, v3 and v4 in both tasks). In each run we incrementally improved the coverage and complexity of the set of heuristics exploited with respect to the previous one:

- **run v1**: initial version of out BARR abbreviation-definition extraction system, including our implementation of the three token-based classifiers of the entity spotting phase (see Section 4.1) and an initial implementation of the relation extraction rules (see Section 4.2);
- **run v3**: with respect to the run v1, we improved the set of relation extraction rules by including heuristics to handle the three special cases of *SF* - *LF* relation listed at the end of Section 4.2 (groups of *SF*s, matching word-initials, *LF* retrieval from the Dictionary of Medical Abbreviations SEDOM). Besides improving the performance of relation extraction, these modifications allowed our system to refine furtherly the set of entities spotted by the three token-based classifiers of the entity spotting phase (see Section 4.1);
- **run v4**: with respect to the run v3, our final run (v4) adds the basic set of heuristics that are tailored to spot cases of nested relations between *SF* and *LF*s, described in the last part of Section 4.2.

In Table 2 and Table 3 we provide the results of the evaluation of our BARR runs, as computed by means of the Markyt Web tool [15]. In particular, Table 2 shows the results

of the entity and relation extraction tasks for each one of our three runs, against the training set of BARR abstracts. We can notice that each new run improves the abbreviation-definition extraction performance.

A consistent evaluation of our abbreviation identification approach against the BARR test set has not been possible due to a bug that affected our system: in our text analysis system we exploited the version 8.4 of the GATE General Architecture for Text Engineering that did not process the texts inside `<![CDATA[ .... ]` sections[4]. As a consequence we were not able to correctly extract abbreviations from a number of abstracts since their text was included in `<![CDATA[ .... ]` sections inside GATE XML documents we used to store the results of intermediate analysis steps. This bug has been identified and solved by releasing, in June 2017, a new version of GATE[5] (version 8.4.1). We realized the presence of this bug when the BARR evaluation period was over, by analyzing the results of our approach over the BARR test set: as a consequence, at the time of writing, we can't provide a bug-free evaluation of our abbreviation identification approach against the BARR test set. Table 3 shows the results of our best run (v4) with respect to the BARR test set. We can notice that, with respect to the performance against the training set (Table 2), the performance of our approach on the test set are considerably lower, probably also due to the bug previously described. Once the BARR test set will publicly released, we plan to consistently evaluate our approach against test data and analyze in details its performance.

BARR task / run	***Precision***	***Recall***	***F1-score***
Entity extraction (run v4)	0.910	0.909	0.909
Entity extraction (run v3)	0.912	0.899	0.901
Entity extraction (run v1)	0.894	0.891	0.893
Relation extraction (run v4)	0.931	0.782	0.850
Relation extraction (run v3)	0.929	0.751	0.830
Relation extraction (run v1)	0.918	0.745	0.822

Table 2. Evaluation of entity and relation extraction performance of the three runs against the training set of BARR abstracts (micro-average, computed by means of Markyt).

BARR task	***Precision***	***Recall***	***F1-score***
Entity extraction	0.722	0.698	0.710
Relation extraction	0.718	0.503	0.592

Table 3. Evaluation of BARR entity and relation extraction tasks (best run, v4) against the test set of BARR abstracts (micro-average, computed by means of Markyt). These results are affected by a bug of the GATE Framework that prevented our system to extract abbreviations from part of the biomedical abstracts included in the BARR test set.

[4] https://sourceforge.net/p/gate/mailman/message/35873689/

[5] https://gate.ac.uk/releases/gate-8.4.1-build5753-ALL/doc/tao/splitch1.html#x4-130001.5.1

6 Conclusions

In this paper we described our participation to the BARR track of IBEREVAL 2107 by introducing our approach to automatically identify abbreviations together with their definitions in Spanish biomedical texts. After a brief introduction of the BARR task, we presented the two main information extraction phases of our system. The first one identifies and characterizes abbreviations and candidate Long Forms by means of a set of token based classifiers. The second phase exploits a collection of heuristics to refine the results of the the first phase and identify relations between abbreviations and Long Forms occurring in the same sentence.

As venue for future research we would like to improve our system by extending and specializing the set of token-level features exploited to automatically extract abbreviations and Long Forms. Moreover we would like to perform more data-based validations and refinement cycles of our relation extraction heuristics. We also plan to evaluate the bug-free version of our approach on the BARR test set, once this data will be publicly released.

References

1. Agerri, R., Bermudez, J., Rigau, G.: Ixa pipeline: Efficient and ready to use multilingual nlp tools. In: LREC. vol. 2014, pp. 3823–3828 (2014)
2. Ao, H., Takagi, T.: Alice: an algorithm to extract abbreviations from medline. Journal of the American Medical Informatics Association 12(5), 576–586 (2005)
3. Chang, J.T., Schütze, H., Altman, R.B.: Creating an online dictionary of abbreviations from medline. Journal of the American Medical Informatics Association 9(6), 612–620 (2002)
4. Chondrogiannis, E., Karanastasis, E., Andronikou, V., Varvarigou, T.: Building a repository for inferring the meaning of abbreviations used in clinical studies. J. Comput 12(1), 76–88 (2017)
5. Cunningham, H., Maynard, D., Bontcheva, K.: Text processing with gate. Gateway Press CA (2011)
6. Gonzalez, G.H., Tahsin, T., Goodale, B.C., Greene, A.C., Greene, C.S.: Recent advances and emerging applications in text and data mining for biomedical discovery. Briefings in bioinformatics 17(1), 33–42 (2015)
7. Hearst, M.S.: A simple algorithm for identifying abbreviation definitions in biomedical text (2003)
8. Intxaurrondo, A., Prez-Prez, M., Prez-Rodrguez, G., Lopez-Martin, J.A., Santamara, J., de la Pea, S., Villegas, M., Akhondi, A., Valencia, A., Loureno, A., Krallinger, M.: The biomedical abbreviation recognition and resolution (barr) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts (2017)
9. Islamaj Doğan, R., Comeau, D.C., Yeganova, L., Wilbur, W.J.: Finding abbreviations in biomedical literature: three bioc-compatible modules and four bioc-formatted corpora. Database 2014, bau044 (2014)
10. Krallinger, M., Intxaurrondo, A., Lopez-Martin, J.A., de la Pea, S., Prez-Prez, M., Prez-Rodrguez, G., Santamara, J., Villegas, M., Akhondi, A., Loureno, A., Valencia, A.: Resources for the extraction of abbreviations and terms in spanish from medical abstracts: the barr corpus, lexical resources and document collection (2017)
11. Liu, Y., Ge, T., Mathews, K.S., Ji, H., McGuinness, D.L.: Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing. pp. 92–97 (2015)

12. Movshovitz-Attias, D., Cohen, W.W.: Alignment-hmm-based extraction of abbreviations from biomedical text. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. pp. 47–55. Association for Computational Linguistics (2012)
13. Nadeau, D., Turney, P.D.: A supervised learning approach to acronym identification. In: Conference of the Canadian Society for Computational Studies of Intelligence. pp. 319–329. Springer (2005)
14. Park, Y., Byrd, R.J.: Hybrid text mining for finding abbreviations and their definitions. In: Proceedings of the 2001 conference on empirical methods in natural language processing. pp. 126–133 (2001)
15. Pérez-Pérez, M., Pérez-Rodríguez, G., Rabal, O., Vazquez, M., Oyarzabal, J., Fdez-Riverola, 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 (2016)
16. Sohn, S., Comeau, D.C., Kim, W., Wilbur, W.J.: Abbreviation definition identification based on automatic precision estimates. BMC bioinformatics 9(1), 402 (2008)
17. Torii, M., Hu, Z.z., Song, M., Wu, C.H., Liu, H.: A comparison study on algorithms of detecting long forms for short forms in biomedical text. BMC bioinformatics 8(9), S5 (2007)
18. Vardakas, K.Z., Tsopanakis, G., Poulopoulou, A., Falagas, M.E.: An analysis of factors contributing to pubmed's growth. Journal of Informetrics 9(3), 592–617 (2015)
19. Vo, T.N.C., Cao, T.H., Ho, T.B.: Abbreviation identification in clinical notes with level-wise feature engineering and supervised learning. In: Pacific Rim Knowledge Acquisition Workshop. pp. 3–17. Springer (2016)
20. Wren, J.D., Garner, H.R., et al.: Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods of information in medicine 41(5), 426–434 (2002)
21. Xu, J., Huang, Y.: Using svm to extract acronyms from text. Soft Computing-A Fusion of Foundations, Methodologies and Applications 11(4), 369–373 (2007)
22. Yamamoto, Y., Yamaguchi, A., Bono, H., Takagi, T.: Allie: a database and a search service of abbreviations and long forms. Database 2011, bar013 (2011)
23. Zhou, W., Torvik, V.I., Smalheiser, N.R.: Adam: another database of abbreviations in medline. Bioinformatics 22(22), 2813–2818 (2006)