Abbreviation Extraction and Normalization in Spanish Clinical Text

Extracción y Normalización de abreviaturas in textos clínicos en español

Areej Mustafa Istaiti
Universidad Carlos III de Madrid
Avda. de la Universidad, 30 28911
100409619@alumnos.uc3m.es

Resumen: En los últimos años se ha producido un incremento en el uso de sistemas de información en salud para explotar la Historia Clínica Electrónica (HCE) como ayuda no solo en los temas administrativos sino también en la gestión y recuperación de información de pacientes bajo distintas perspectivas (práctica e investigación clínica, atención médica, etc.). Ello requiere el desarrollo de técnicas de procesamiento automático para obtener información de manera ágil convirtiendo información no estructurada en información no estructurada y procesable por algoritmos en procesos de toma de decisiones. En esta tesis, subrayamos la importancia de trabajar con terminología biomédica que ayude a la comprensión de narrativa clínica, particularmente las abreviaturas. Se presentan algunos de los trabajos del estado de la cuestión para reconocer y desambiguar abreviaturas, la propuesta de investigación para textos clínicos en español que ha recibido poca atención, así como los retos pendientes en este campo.
Palabras clave: extracción de información, narrativa clínica, abreviaturas, reconocimiento de entidades.

Abstract: In recent years, there has been an increase in computerized healthcare systems and the accompanying use of electronic records to facilitate patients' administrative issues as well data management and information retrieval from different perspectives (medical care, clinical research, etc.). This requires the development of automatic techniques to obtain information in a more agile way, making unstructured information structured and actionable by algorithms, thus facilitating strategic decision-making. In this research, we highlight the importance of working with biomedical terminology to understand clinical narrative, particularly concerning abbreviations. Some of the state-of-the-art solutions to recognize and resolve them, our research proposal for Spanish clinical text that has hardly been investigated as well as the open challenges in this field are also introduced.
Keywords: information extraction, clinical text, abbreviations, named entities recognition,

1 Introduction

With the evolution of the computerized system, and the development in the health care sector, many terms have been accompanied by these systems; such as Electronic Medical Record (EMR), Electronic Health Record (EHR) or Computer Based Record (CPR). Despite the differentiate of the terminology, the goal of using this is to keep information about patients medical and treatment history concerning a patient in the national health system, including demographic information, diagnoses, laboratory tests and results, prescriptions, radiological images, clinical notes, and more (Birkhead, Klompas, and Shah 2015).

Two forms of data stored in electronic records; structured data which followed the predefined model and rules to store. Hence it is accurate data with no specific errors (called metadata). The second form is unstructured such as free text, X-rays images, scanned files which are not formatted and also need to be considered.

Using these records in healthcare and clinical research requires transforming unstructured data into structured data that could be input to algorithms to solve medical

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 13–19
Bilbao, Spain, September 25th 2019.

problems and support clinical decisions. Apart from exploiting this information, this structuration could facilitate exchanging of data between different hospitals and primary care centers known as “semantic interoperability,” for instance, normalizing terminologies that make vocabularies a shared meaning among various organizations. Besides, working in health-related documents readability (such as discharge summary reports of patients) where complex terms have to be simplified is another area of interest.

Currently, only structured metadata is processed. The rest of the information, in an unstructured format (free text, images, video), remains without being able to be exploited by automatic processes. Approximately 80% of clinical data are not structured, and consequently cannot be used by algorithms and contribute to decision making. The development of technology capable of processing and exploiting unstructured information in clinical text from electronic records in the current context of big data can have many applications, both in improving clinical practice (automatic generation of summaries of episodes related to a patient or group of patients, clinical decision support systems to customize diagnoses and treatment of diseases, infectious disease alerts, etc.) and research (semi-automation of epidemiological studies, for example in the identification of patient cohorts).

Transforming clinical narrative in structured controlled vocabulary is a challenge for several reasons; apart of the complexity of extracting relevant facts from free text, in the case of clinical text, it is susceptible to spelling errors, ungrammatical sentences and containing a large number of medical abbreviations because it is speedy written under pressure of work, and it is rarely checked before to store it.

Abbreviations are universal phenomena, occurring in all languages and writings, and it could be formed in several ways. Table 1 shows how the abbreviation could be formed (Zahariev 2004).

Abbreviations are a particular type of biomedical named entities, and currently named entity recognition (NER) techniques could be used/adapted to work with this specific terminology. An abbreviation is a short form of a word or a phrase. For instance, ´NKB´ is an abbreviation of "nuclear factor-kappa B". The abbreviation is called short form (SF), and the definition or expansion of abbreviation is called long form (LF). As a result of NER process over a text, a list of disambiguated <SF, LF> pairs should be obtained.

Abbreviation form	**SF**	**LF**
Truncating the end of LF	adm	**adm**inistration
First letter initialization from each word	AAA	**a**bdominal **a**ortic **a**neurysm
Syllabic initialization	BZD	**b**en**zo**d**i**azepine
combination of the beginning of some of the words of LF	ad lib	**ad lib**itum
Symbols/synonyms substitution or initialization	ASD I	**Primum a**trial **s**eptal **d**efect

Table 1：Examples of how abbreviations are formed

There are many knowledge sources available in the biomedical domain which contains abbreviations and its long forms such as the unified medical language system (UMLS), AcroMed (Pustejovsky et al. 2001) and SaRAD (Adar 2004) but there is still no comprehensive list of the abbreviations, and each database has its definition schema. Also, there are many NLP tools like cTAKES system (Chute et al. 2010), MetaMap (Aronson 2001) and MedLee (Friedman et al. 1994) which could be used in abbreviations extractions process.

2 Research problem

In the medical text, there are many classifications for abbreviations types. According to (Birkhead, Klompas, and Shah 2015) mention two types of abbreviations related to the appearance of its long form, global if it appears in the text without their definition, local if it appears with their definition in the same text. While (Yu, Hripcsak, and Friedman 2002) presents another classification from a different view, dynamic and common abbreviations are distinguished. A dynamic abbreviation is valid for particular articles. In contrast, a common abbreviation is accepted to be as synonyms in their domains.

Other issues that have to be considered are that there are no rules for the creation of

abbreviations as mentioned in table 1 previously; also it could contain special characters or numbers for example "alanina amino transferasa (A.L.T.)". But not necessarily the world with special characters be an abbreviation like "Ph.D.". The fast creation of these terms (in Medline abstracts, approximately there are 65.000 new abbreviations in 2004). The scope of the abbreviation (an abbreviation could be established, and normalized term contained in standardized resources or could have the scope of a hospital or even a healthcare professional). The occurrences of multilingual <SF, LF> pair (for instance, ´OCT´ is "Optical coherence tomography" and is used in Spanish clinical texts although the equivalent expansion in Spanish is "tomografía óptica de coherencia " and the corresponding abbreviation should be ´TOC´). However, the most critical problem is related to ambiguity, a high percentage of abbreviations have several expansions or LF; for example, 'ABC' could be "Antigen-Binding Capacity" and "Advanced Breast Cancer" and this could lead to a severe problem if it is incorrectly identified.

Medline is a database that stores articles of the biomedical domain. In particular, 80% of the abbreviations defined in the unified medical language system (UMLS) have ambiguous occurrences in MEDLINE (Liu, Lussier, and Friedman 2001). Besides, there is no standard benchmark to evaluate these approaches since each tool builds its corpus to test the performance.

Furthermore, with the abundance of biomedical abbreviations databases and tools, there is still no complete list of existence abbreviations, this due to quick creation for it.

The Spanish language is considered the second spoken language over the world and most of the algorithms used to extract the abbreviation works for the English language, since the Spanish language has its specification and differs from the English language there is a need to implement an algorithm that deals with Spanish medical documents.

3 Background and related work

The extraction process consists of detecting <SF> candidates from medical documents firstly, then detect <LF> candidates if they are mentioned in the same text and lastly maps the most suitable <LF> to the adequate <SF>. Several approaches are found to implement these steps: (I) alignment algorithm approach, (II) pattern matching approach, (III) statistical approach, and (IV) machine learning approach. The following paragraphs show these approaches in detail.

(Schwartz and Hearst 2003) introduces an alignment algorithm based on the assumption that both the SF and LF appear in the same text. The SF at least must have two letters and the candidate long form should have no more than min (|A| + 5, |A| * 2) words, where |A| is the number of characters in the short form. Then a backward strategy begins to map the long form with the most suitable long form. Taking into their account that a letter in the short form could be an interior letter in the long form. They achieved 96% precision and 82% recall on the Medstract (Pustejovsky et al. 2001) corpus.

Another approach is followed to extract the abbreviations is a rule-based approach which SF candidates are found based on a set of rules and punctuation. Then many LF candidates are gathered from nearby words that appear around the SF. The SF and the LF are connected by rules, such as occurrences and order of short form letters in long form using a specific stop word list for reducing potential errors in the output.

(Yu, Hripcsak, and Friedman 2002) applied this approach to map both defined and undefined abbreviations to their full form. Yu considered defined abbreviations, and their full form could appear in two different forms <LF><(SF)> or <SF><(LF)>, then he applied pattern matching rules to find the right long form for the SF candidate. For undefined abbreviations, he used different databases as (Genbank, LocusLink, LRABR) to map it with LF. The system was evaluated on 50 articles from medical and biological domains and achieved 70% recall and 95% precision.

This approach is easy to implement and readable for human. On the other hand, domain experts are needed to build a set of rules in a precise way. However, the main drawback is the construction of rules dealing with hundreds of cases that make the process is tedious.

The third approach which could be followed is a machine learning approach. The model which follows this approach is trained using an annotated data set (labeled data) firstly; after that, this classifier is used to predict the new data. (Chang, Schütze, and Altman 2002) Uses

dynamic programming to detect if there is a possible alignment between the abbreviation and its expansion, and the result is fed to compute feature vectors for identifying the correct expansion. He applies linear regression on a pre-selected set of features. Also, the algorithm was evaluated on Medstract corpus; the recall/precision was 95% at 75%. In general, machine learning based approaches depend on the learning model and the training data and require much labor and a long time preparing the training set.

Finally, statistical approaches usually concentrate on extracting abbreviations that frequently are used in biomedical text, and it needs a large dataset. (Okazaki and Ananiadou 2006) used statistical methods depending on co-occurrences for LF-SF achieving 99% precision and 82–95% recall on evaluation corpus that roughly emulates the whole MEDLINE. This type of approaches needs a long time to do the statistical methods. Table 2 summarizes the current approaches with recall and precision figures.

For the Spanish language, the work still on its first stages, IberEval has been held in 2017 and 2018, consequently. This kind of challenges aims to support the development of Human Language Technologies (HLT) for Iberian languages (Spanish, Portuguese, Catalan, Basque and Galician), by creating series of evaluation and a discussion forum about Natural Language Processing systems on an ongoing basis (https://sites.google.com/view/ibereval-2018)The challenge in its 2018 version involved Biomedical Abbreviation Recognition and Resolution track (BARR2) (http://temu.bsc.es/BARR2/)This track composed of two tasks, the first one is for local abbreviations detections and the second for ambiguity problems.

Three participants collaborated with this task different approaches were used to accomplish the task goal; Vicomtech (Cuadros et al. 2018) system which extracts SF candidates based on different machine learning algorithms and heuristic based, then check the LF into their dictionary, if the LF doesn't exist they applied a heuristic rule to extract the LF in eighth n=gram surrounding the SF. The best precision result the system got is 88.56% in a combined machine learning and regular expression. Recall 76.05% and f-measure 81.71% when the three approaches were combined.

MAMTRA-MED (Montalvo et al. 2018) system is a combination of a pattern-based and dictionary-based approach. They detect terms in capital letters or combinations of capital letters with lower case letters, numbers, and other characters. The best precision was 91.20% for the dictionary-based system and recall 73.53% f-measure 79.01% when the system prioritizes the relations found by the dictionary-based.

ARBRex (Sánchez-León 2018) uses a pattern match approach for creating a dynamic regular expression to detect SF and LF. The system was evaluated and got a precision 88.61 %, recall 88.23%, and f-measure 88.42%.

Approach	System	Recall	Precision
Alignment Algorithm	(Schwartz and Hearst 2003)	82%	96%
Rule-based	(Yu, Hripcsak, and Friedman 2002)	70%	95%
Statistical	(Okazaki and Ananiadou 2006)	99%	82%
Machine learning	(Chang, Schütze, and Altman 2002)	95%	75%

Table 2: Current approaches for abbreviation extraction and evaluations of the systems

4 Proposed work

The research work is defined around two important objectives. Firstly, a schema including relevant information for biomedical abbreviations should be defined that allows us to integrate existing repositories and gazetteers like UMLS, ADAM, Acromed, and SNOMED-CT. This is essential to overcome the problem of coverage of these databases as well as to keep relevant information about provenance, language, composition among others. Secondly, a robust approach to recognize an disambiguate abbreviations I Spanish biomedical text. A hybrid approach combining knowledge based and machine learning could be an interesting starting point.

Figure 1 represents these objectives distinguishing back end and front end sides of an architecture to face the problem of working with abbreviations. A system be used to detect SF candidates from the biomedical text and

then classify it as a valid one or not using different approaches (pattern or rule based, machine learning). After getting a list of valid abbreviations, LF candidates will be detected for each SF, then mapping SF for its LF in the same text (local abbreviation). In this training phase, pre-processing steps will be applied to read the sentences separately, tokenize it, and after building a model using a training data, the second phase will be testing the model exclude the special words (as seen in Figure 1)using a dataset. For ambiguity problem and global abbreviations, neural network models will be explored trying to exploit the common abbreviations repository which is built in the back-end phase.

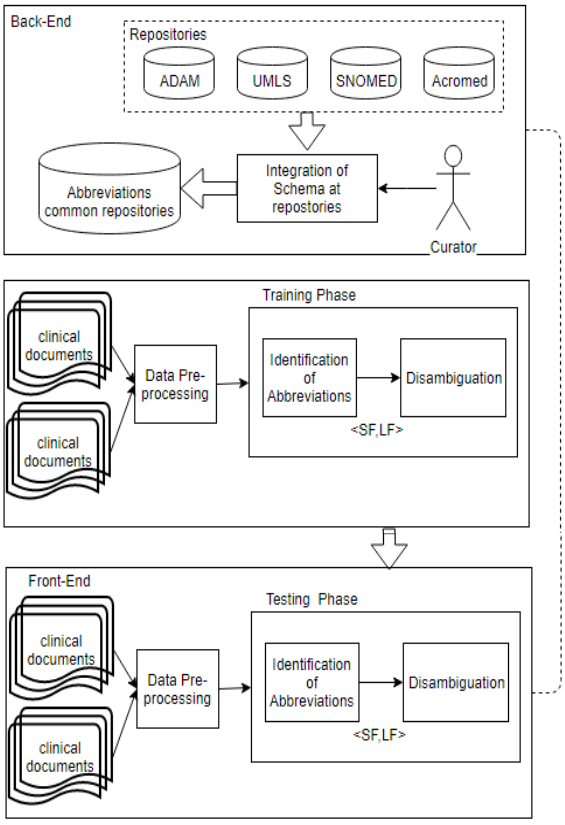

Figure 1: Framework to recognize abbreviations in clinical texts

5 Preliminary method

5.1 Dataset

BARR corpus was used, as a first step of the work. It contains 3563 clinical reports gathered from Medline database, Spanish Bibliographic Index in Health Sciences (IBECS), and Scientific Electronic Library Online (SciELO). See Figure 2 and Figure 3 with an example extracted from BARR dataset .

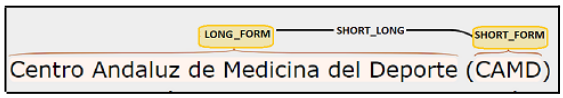

Figure 2: BRAT screenshot for an abbreviation S1888-75462014000200009-1.txt

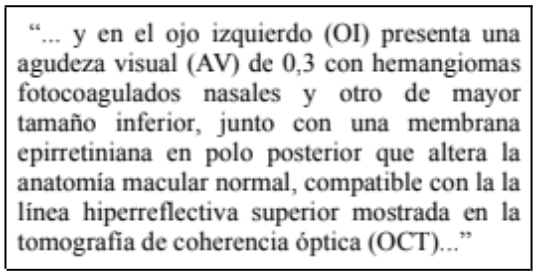
"... y en el ojo izquierdo (OI) presenta una agudeza visual (AV) de 0,3 con hemangiomas fotocoagulados nasales y otro de mayor tamaño inferior, junto con una membrana epirretiniana en polo posterior que altera la anatomía macular normal, compatible con la la línea hiperreflectiva superior mostrada en la tomografía de coherencia óptica (OCT)..."

Figure 3: sample of BARR2 Corpus from S0004-06142008000100008-1.txt file

5.2 Method

Subtask 1 of BARR track was chosen to run to achieve the initial goal. This task was about the detection of explicit occurrences of abbreviation-definition pair that found in their annotated corpus.

(Schwartz and Hearst 2003) algorithm was used to be executed on BARR corpus; this algorithm, as mentioned in the related work section, is based on an alignment approach and it uses several patterns to recognize <SF,LF> pairs. Table 3 below shows abbreviations that detected by the algorithms based on the pattern which they used.

Condition	True Positive
Consist of at most two words	AO
Their length is between two to ten characters	ASLO
At least one of these characters is a letter	angio-TC
The first character is alphanumeric,	U.I.

Table 3: Example of True positive abbreviation detected by (Schwartz and Hearst 2003)

5.3 Result

In this first experiment, 135 abbreviations were detected in total, 15 are considered as wrong abbreviations, 32 abbreviations were missed, with precision 88%, recall 67%, and F-measure 76%. Table 4 shows some examples of undetected abbreviations (false negatives) from BARR corpus.

Type of errors	False negative
Long form includes additional words	Hidratos de Carbono (HC)
Skipped characters in the SF	cadenas ligeras kappa (CLL-K)
Out of order LF	Transaminasa glutámico-pirúvica (GPT)
One-character SF	temperatura (T)
SF doesn't exist between parentheses	(fracción de eyección, FE: 0,61)

Table 4: sample cases which Schwartz and Hearst algorithm did not detect.

Acknowledgment

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R).

References

Adar, Eytan. 2004. "SaRAD: A Simple and Robust Abbreviation Dictionary." *Bioinformatics* 20(4): 527–33.

Aronson, A R. 2001. "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program." *Proceedings. AMIA Symposium*: 17–21. http://www.ncbi.nlm.nih.gov/pubmed/11825149%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2243666.

Castaño, José et al., 2018. "A Simple Approach to Abbreviation Resolution at BARR2, IberEval 2018." *CEUR Workshop Proceedings* 2150: 316–21.

Chang, Jeffrey T., Hinrich Schütze, and Russ B. Altman. 2002. "Creating an Online Dictionary of Abbreviations from MEDLINE." *Journal of the American Medical Informatics Association* 9(6): 612–20.

Chute, Christopher G et al. 2010. "Mayo Clinical Text Analysis and Knowledge Extraction System (CTAKES): Architecture, Component Evaluation and Applications." *Journal of the American Medical Informatics Association* 17(5): 507–13.

Cuadros, Montse, Naiara Pérez, Iker Montoya, and Aitor García Pablos. 2018. "Vicomtech at BARR2: Detecting Biomedical Abbreviations with ML Methods and Dictionary-Based Heuristics." *CEUR Workshop Proceedings* 2150: 322–28.

Friedman, Carol et al. 1994. "A General Natural-Language Text Processor for Clinical Radiology." *Journal of the American Medical Informatics Association* 1(2): 161–74.

Gaudan, S., H. Kirsch, and D. Rebholz-Schuhmann. 2005. "Resolving Abbreviations to Their Senses in Medline." *Bioinformatics* 21(18): 3658–64.

Liu, H, Y A Lussier, and C Friedman. 2001. "A Study of Abbreviations in the UMLS." *Proceedings. AMIA Symposium*: 393–97. http://www.ncbi.nlm.nih.gov/pubmed/11825217%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2243414.

Montalvo, S., R. Mart, M. Almagro, and S. Lorenzo. 2018. "MAMTRA-MED at Biomedical Abbreviation Recognition and Resolution - IberEval 2018." *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*: 290–96.

Okazaki, Naoaki, and Sophia Ananiadou. 2006. "Building an Abbreviation Dictionary Using a Term Recognition Approach."

Bioinformatics 22(24): 3089–95.

Pustejovsky, James et al. 2001. “Automatic Extraction of Acronym-Meaning Pairs from MEDLINE Databases.” *Studies in Health Technology and Informatics* 84: 371–75.

Sánchez-León, Fernando. 2018. “ARBOREx: Abbreviation Resolution Based on Regular Expressions for BARR2?” *CEUR Workshop Proceedings* 2150: 303–15.

Schwartz, Ariel S., and Marti A Hearst. 2003. “A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text.” *Symposium A Quarterly Journal In Modern Foreign Literatures* 462: 451–62.

Yu, Hong, George Hripcsak, and Carol Friedman. 2002. “Mapping Abbreviations to Full Forms in Biomedical Articles.” *Journal of the American Medical Informatics Association* 9(3): 262–72.

Zahariev, Manuel. 2004. “8. Acronyms.” *COSPAR Information Bulletin* 1999(144): 50–51.