Abbreviation Extraction in Spanish Clinical Text

Areej Istaiti , Paloma Martínez
Department of Computer Science and Engineering
Universidad Carlos III de Madrid
Madrid, Spain

ABSTRACT

In recent years, there has been an increase in computerized health-care systems and the accompanying use of electronic records which mainly is found to improve health-care performance. Storing patient's information electronically in a free text format raises some difficulties. This work highlights the importance of working in extraction information from Electronic Medical Records to understand clinical narrative, particularly concerning abbreviations. Our research proposal focuses on Spanish clinical text.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Information extraction**; **Machine learning**;

KEYWORDS

EMR, NLP, Abbreviation, SF, LF

1 INTRODUCTION

Electronic Medical Records (EMR) keep medical and treatment history concerning a patient in the national health system. EMR unstructured format data (free text, images, video, ...) remains without being able to be exploited by automatic processes. Approximately, 80% of clinical data are unstructured, and consequently can not be used by algorithms and contribute to decision making. Abbreviation is defined as a short form of a word or a phrase which are used frequently in EMR, and it is considered biomedical named entities. For instance, NKB is known as short form SF and "nuclear factor-kappa B"its Long Form LF. Dealing with abbreviations is one of the most critical challenges in the medical field because failure to understand them could lead to medical misunderstandings and other implications. Named Entity Recognition (NER) techniques can be used to work with this special terminology for finding the most suitable LF for the SF, and as a result of this process over a text, a list of disambiguated <SF, LF> pairs could be obtained.

2 RESEARCH PROBLEM

There have been more than 197,000 unique medical abbreviations found in the clinical text [2], each SF could have more than LF due to its random formations and depending on the scope it is used in. Also, the occurrences of multilingual <SF, LF> pair (for instance, PSA is "Prostate Specific Antigen" and is used in Spanish clinical texts although the equivalent expansion in Spanish is "Antígeno Prostático Específico" and the corresponding abbreviation should be APE). These issues represent a severe problem of dealing with abbreviations, moreover 80% of abbreviations that were found in (UMLS) are ambiguous [1]. Natural Language Processing (NLP) is used to solve this kind of problems. Most of the relevant work focuses on the English language to solve the problem, hence, the

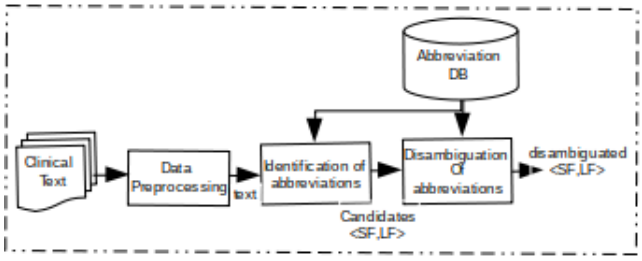

Figure 1: Proposed Abbreviation Extraction Architecture.

Spanish language is considered the second spoken language over the world and Spanish language has its specification that differs from the English language, there is a need to implement a system that deals with Spanish medical documents. Many approaches are used to enhance dealing with medical abbreviations[3]. The rule-based approach is better for a simple set of data because it needs a lot of patterns to match abbreviations. The statistical approach depends on a large amount of data which makes the executions time is too long. And machine learning approach gives the best result for this process; training data could be used to build the model then test data set to evaluate it.

3 PROPOSED SYSTEM

A hybrid approach will be applied in Spanish clinical text; at the first step, the text will be tokenized, normalized and substituted for processing. Then a pattern matching approach will be applied to extract the <SF, LF> candidates from the text, and machine learning algorithms will also be applied to map the most suitable LF for every SF. The ambiguity of abbreviations will be taken into account and mapping to common abbreviations database. Figure 1 shows the proposed architecture.

ACKNOWLEDGMENTS

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

[1] H Liu, Y A Lussier, and C Friedman. 2001. A study of abbreviations in the UMLS. *Proceedings. AMIA Symposium* (2001), 393–7. http://www.ncbi.nlm.nih.gov/pubmed/11825217

[2] Yue Liu, Tao Ge, Kusum S. Mathews, Heng Ji, and Deborah L. McGuinness. 2015. Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion. (2015), 1–6. https://doi.org/10.18653/v1/W15-3810 arXiv:1804.04225

[3] Manabu Torii, Zhang Zhi Hu, Min Song, Cathy H. Wu, and Hongfang Liu. 2007. A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics* 8, SUPPL. 9 (2007), 1–9. https://doi.org/10.1186/1471-2105-8-S9-S5