The use of Natural Language Processing techniques to support Health Literacy: an evidence-based review

Paloma Moreda, Elena Lloret

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain

Abstract

Background and objectives: To conduct a literature search and analysis of the existing research using natural language processing for improving or helping health literacy, as well as to discuss the importance and potentials of addressing both fields in a joint manner. This review targets researchers who are unfamiliar with natural language processing in the field of health literacy, and in general, any researcher, regardless of his or her background, interested in multi-disciplinary research involving technology and health care.

Methods: We introduce the concepts of health literacy and natural language processing. Then, a thorough search is performed using relevant databases and well-defined criteria. We review the existing literature addressing these topics, both in an independent and joint manner, and provide an overview of the state of the art using natural language processing in health literacy. We additionally discuss how the different issues in health literacy that are related to the comprehension of specialised health texts can be improved using natural language processing techniques, and the challenges involved in these processes.

Results: The search process yielded 235 potential relevant references, 49 of which fully fulfilled the established search criteria, and therefore they were later analysed in more detail. These articles were clustered into groups with respect to their purpose, and most of them were focused on the development of specific natural language processing modules, such as question answering, information retrieval, text simplification or natural language generation in order to facilitate the understanding of health information.

Keywords: Health Literacy, Health Informatics, Medical Informatics, Natural Language Processing, Information Technology

1. Introduction

Literacy is a fundamental human right and the foundation for lifelong learning. Definitions and understandings of this concept have broadened considerably over the past fifty years into two points of views. On the one hand, literacy can be understood as the set of tangible skills, especially those cognitive skills of reading and writing that are independent of the context in which they are acquired and the background of the person who
acquires them [1]. On the other hand, the National Assessment of Adult Literacy (NAAL) [2] considers literacy as both task-based and skills-based. The task-based definition of literacy focuses on the everyday literacy tasks an adult can and cannot perform; whereas the skills-based focuses on the knowledge and skills an adult must possess in order to perform these tasks. Such skills range from word-level skills (e.g., recognising words) to higher level skills (e.g., making inferences from text). In both cases, a central and important issue in literacy is the text understanding.

In recent years, the concept of literacy has been also applied to specific contexts, such as information, visual, scientific, or health, among others. In particular, health literacy is the ability to understand and act on health information [3]. This includes the understanding of health concepts, and the terminology used by doctors and nurses (medical jargon). In medical and health documents, a lot of specialised terminology and vocabulary is employed, and the texts are written following a particular style. As a consequence, this is often hard to understand by non-experts in the medical domain, even if the user is literate or with a perfect reading comprehension ability. The inability to understand this type of information can lead to problems, such as low interaction between doctors and patients, or incorrect administration of a treatment. According to the European project IROHLA [4], people with sufficient health literacy skills are able to act proactively on health issues, take their own decisions and manage health and illnesses well. This project states that in EU countries, 10%-30% of the population has insufficient health literacy skills, which is associated with higher morbidity and mortality, while utilisation of health services is higher and treatment outcomes are poorer than average. Therefore, health literacy is undoubtedly an important issue in our society.

Moreover, there is another problem concerning the exponential increase of health information on the Internet. In our current digital information society, it is very common that users search specific information about diseases, symptoms and treatments in fora, blogs, or general health Web pages, where the language used is much simpler and understandable, since most of this information may be written by other non-expert users that suffered from the same disease, received similar treatments, etc. However, a potential associated risk of this concerns the fact that not all users check the source where the information comes from, thus leading to problems, such as the reliability of the information, that could not have been provided or supported by health professionals. This issue makes the role of health literacy more important in our society. Ideally, it would be necessary either to make users aware of the relevance of this topic for their daily lives, and therefore provide them with at least basic training, or encourage health experts to make the information more accessible to patients. Both issues are not easy to address, so why not exploiting the use of technology to analyse whether it would be possible to support health literacy to some extent, thus avoiding the costly task to adapt the information in a manual way?

Specifically, Natural Language Processing (NLP) is a field of artificial intelligence that, on the one hand, investigates the properties of written human language to model the cognitive mechanisms underlying the under-
Standing and production of written language and, on the other hand, develops applications involving the intelligent processing of human language by computers [5]. NLP can, for example, help users find relevant information according to their interests, perform automatic text simplification for adapting a text to the user knowledge level, or determining the sentiment a user has towards an issue. These are just some of the multiple possibilities of NLP. For our aims, text simplification is of special interest because this task facilitates reading comprehension to people with specific reading disabilities, such as down’s syndrome [6], dyslexia [7], or autism [8]. Moreover, it has an enormous potential for other purposes, not only the ones related to reading comprehension problems. For instance, the advantages of using NLP in the field of health literacy are that potential lexical, syntactic and semantic obstacles appearing in the texts could be automatically identified and the text could be simplified and enriched through the use of illustrative images, definitions, and/or synonyms.

There is no doubt that mechanisms and tools to improve and/or facilitate the understanding of medical texts would be very useful and relevant in the present society. These type of tools would help users to access and comprehend specialised health documents, without requesting explicitly health professional to provide such information in an easier manner. As a consequence, the improvement of health literacy will prevent users from searching and consuming medical information on the Web (blogs, fora, etc.), where the content may have not been necessarily written and/or validated by health professionals. It is in this context where NLP techniques can be very useful for making medical and health information more accessible.

1.1. Context and illustrative scenario

Previous research works have addressed the issue of health literacy with the aim of providing more accessible information, and different techniques for improving the accessibility of materials created for patients have been suggested. These techniques comprise the use of simple and common words; short sentences; or writing in the active voice, to name just a few [9, 10, 11, 12, 13]. However, in order to follow these guidelines and recommendations, medical professionals have to be aware of the problematic obstacles, and try to explain or write them in a simpler way. This may be not always feasible, since clinicians often overestimate the ability of patients to understand medical information, and therefore they do not see the need for creating text that is suitable for each type of patient [14]. So, it is here where NLP plays an important role, helping professionals to do this task in an automatic manner.

There has been extensive research into developing NLP tools for the medical domain [15, 16, 17, 18]. To the best of our knowledge, most of these proposals are preliminary studies about how NLP could be used in this medical domain and they are focused on the clinicians’ perspective rather on the patient one. However, the current developed NLP tools and resources can be employed for improving the reading comprehension of medical information, thus reducing health literacy based disparities in our society [18]. Using the appropriate NLP resources, specialised medical terminology could
be automatically detected and highlighted, providing in addition, its definition and synonyms. For instance, if we consider the following fragment of text, extracted from a drugs’ patient information leaflet\(^1\), we would find a number of terms that may be difficult to understand, either for being specialised terminology (e.g., *dysphoric*), or for not being frequently-used terms (e.g., cessation or restlessness):

*Smoking cessation* with or without treatment is associated with various symptoms. For example, *dysphoric* or *depressed* mood; *insomnia*, *irritability*, *frustration* or anger; *anxiety*; difficulty concentrating; *restlessness*; decreased heart rate; increased *appetite* or weight gain have been reported in patients attempting to stop smoking. No attempt has been made in either the design or the analysis of the CHAMPIX studies to distinguish between adverse events associated with study drug treatment or those possibly associated with *nicotine withdrawal*.

Using some of the existing technology, for example analysing this paragraph with the NLP techniques developed within the FIRST project\(^{[19]}\), several concepts that may be difficult to understand by users could be automatically detected (i.e., the ones shown in boldface). These concepts belong to the specific terminology of the medical domain, such as *dysphoric* (feeling unhappy), or *insomnia* (not being able to sleep). As it was previously stated, besides specialised vocabulary, NLP techniques could also detect other types of words, e.g., infrequent terms, such as cessation, appetite, restlessness, or withdrawal. Users may not familiarised with them either, and therefore, the text could be difficult to understand. However, through the use of NLP technology, the complexity of the text could be automatically reduced, by enriching the concepts with more common synonyms and/or definitions, such as *stopping*, *desire for food*, *excitement*, or *removal*, respectively.

This type of processes are now possible due to the availability of NLP resources that are matured enough, and have been developed to bridge the gap between NLP and medicine. All the information and knowledge that NLP tools and resources provide can be used for including assistive elements, detecting and resolving obstacles, and/or personalising a text. This will allow that a complex specialised text becomes easier to understand by anybody, regardless his/her expertise in the health field. Examples of NLP techniques and resources that can be used to make a text more accessible for any user in an automatic manner include Word Sense Disambiguation, a NLP task the aim of which is to correctly determine the sense of a word in a context; WordNet ontology\(^{[20]}\), a general-purpose lexical-semantic resource containing words, their definitions and synonyms, as well as the different types of relationships with other words in the ontology; or WordNet Domains\(^{[21]}\), which groups the common concepts according to different domains, including medicine. In addition to the previous NLP resources, others that have been specifically developed for the medical domain are also available. This is the case of Unified Medical Language System (UMLS), which is a resource

\(^1\)http://www.medicines.org.uk/emc/document.aspx?documentid=19045 (last access 21 May 2015)
that groups health and biomedical vocabularies and standards to enable interopera- 
tibility between computer systems [22]; or Health Information Text Extraction (HITEX) system\textsuperscript{2}, which can extract specific information, (e.g., diagnosis extraction, and discharge medications) [23]. Thanks to these resources (e.g. UMLS), drug names, such as CHAMPIX, could have been also detected in the aforementioned example text fragment.

1.2. Goal and structure

Under the above-mentioned premises, the main objective of this article is to review and analyse to what extent the technology, and in particular, NLP has been employed and applied to support health literacy in some way. If so, we would like to also study for which purpose and how NLP was used, and if not, which other types of technology, when available, were employed. In order to achieve our objective, we conducted an evidence-based review within three of the most relevant bibliographic databases: PubMed, Scopus, and Cochrane, analysing the existing research works in detail, and drawing interesting conclusions.

After this introduction, the remaining of this article is as follows. Section 2 contains the methodology employed for conducting the evidence-based review. Section 3 provides the results obtained from different perspective, together with discussing also the most important research works. Further on, Section 4 we analyse the potentials and implications that NLP may have for the future of health literacy, and finally, Section 5, concludes the article by summarising its main ideas.

2. Methods

This section explains the methodology used for conducting a evidence-based review of the research addressing the use of technology, and in particular NLP for health literacy. We followed the STARLITE methodology [24], which takes into consideration the essential elements for reporting literature searches (sampling strategy, type of study, approaches, range of years, limits, inclusion and exclusions, terms used, electronic sources).

2.1. Sampling strategy

In order to collect a sample of research articles to analyse, we will used the browsers of different databases to obtain the title, abstract and full content of the articles that match specific search criteria that will be later outlined. For performing the search of relevant literature, we will rely on MeSH descriptors independently whenever possible (if not we will considered MeSH descriptors as keywords), their synonyms, as well as the combination of them using boolean operators (e.g., "and", "or", etc.).

MeSH is the National Library of Medicine’s controlled vocabulary thesaurus\textsuperscript{3}. It is a thesaurus consisting of sets of terms naming descriptors in a twelve-level hierarchical structure that permits searching at various levels.

\textsuperscript{2}https://www.i2b2.org/software/projects/hitex/hitex_manual.html
\textsuperscript{3}http://www.nlm.nih.gov/mesh/MBrowser.html
of specificity, and it is used for indexing articles. Each bibliographic reference is associated with a set of MeSH terms that describe the content of the item. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. Broad headings can be found at the most general level of the hierarchical structure (e.g. Anatomy), whereas more specific headings are found at narrower levels (e.g. Ankle). In total, there are 27,149 descriptors in 2014 MeSH. There are also over 218,000 entry terms that assist in finding the most appropriate MeSH Heading.

Moreover, we establish the criterion that if the same research work appears in more than one database, it will be only analysed once.

2.2. Type of study
For our study, we were interested in looking information appearing in research articles, systematic reviews, surveys, chapter in books or books. Editorials, notes, and other type of publications were of no interest for our analysis.

2.3. Approaches
Several search strategies are defined using the MeSH descriptors, and then a manual inspection and analysis over the most relevant publications with respect to the purpose of this article were conducted. First, an initial inspection of the title and abstract was carried out to select or discard references, and further one, the whole content of the article was analysed.

2.4. Range of years
Initially, we did not established any specific period for the search, since we are more interested in knowing when health literacy started to be an important topic, and how the technology was applied.

2.5. Limits
Our evidence-base search will be limited to articles in English and Spanish.

2.6. Inclusion and exclusions
A set of inclusion criteria were defined for accepting the retrieved articles for the analysis, which are shown in Table 1. The articles that do not meet such criteria were discarded from the analysis.

The reason for applying the restriction for the retrieved results to be only in English or Spanish is due to the context of influence, where Spanish language is at the immediate context, and thus, we are interested in knowing what are the advances and developments carried out in Spanish, and English at the global, since it is the language in which most research is published. The remaining criteria are focused on the topics we are interested in and therefore, they will lead to more precise searches.
Table 1: Inclusion criteria for the literature search.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Research publications written in English or Spanish.</td>
</tr>
<tr>
<td>- Research aimed at analysing and/or using artificial intelligence, including</td>
</tr>
<tr>
<td>NLP for supporting health education, including consumer health education and health</td>
</tr>
<tr>
<td>literacy (both topics in conjunction).</td>
</tr>
<tr>
<td>- Research in which the topics of Natural Language Processing and/or Health Literacy</td>
</tr>
<tr>
<td>are addressed as main topics.</td>
</tr>
</tbody>
</table>

2.7. Terms used

We accessed MeSH browser⁴ and inspected it looking for the best descriptors concerning our topics of interest, trying to be as precise as possible. In this manner, taking into account that our main aim was to focus on the analysis of how NLP was employed for health literacy, our search terms were taken from the MeSH descriptors are shown in Table 2.

Table 2: Descriptors and synonyms used for the search.

<table>
<thead>
<tr>
<th>MeSH descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Information Science; Artificial Intelligence;</td>
</tr>
<tr>
<td>Natural Language Processing; AI; Machine Intelligence</td>
</tr>
<tr>
<td>- Health Education; Consumer Health Information; Health Literacy</td>
</tr>
</tbody>
</table>

After analysing all possible MeSH descriptors, for our search strategy we finally selected the descriptors that exactly matched with our topic keywords (ie., Natural Language Processing and Health Literacy), and consequently we searched for literature containing either one of these descriptors or both of them using the boolean operator and. However, in other to broaden the search, we also took into account the top 3 descriptors that were above the selected ones in the MeSH hierarchy. In this manner, for Natural Language Processing, we also search for literature concerning Information Sciences, and Artificial Intelligence; and for Health Literacy, the descriptors Consumer Health Information and Health Education were also considered. Having a look at the synonyms of the descriptors, we also performed searches with the descriptors AI and Machine Intelligence, that were synonyms for Artificial Intelligence.

2.8. Electronic Sources

Three databases were used: PubMed\(^5\), Cochrane\(^6\), and Scopus\(^7\). The reasons for choosing these databases among all the existing ones are next provided. First, we wanted that our search included databases containing articles not only of the health and medical domain, but also, dealing with artificial intelligence applied to the previously mentioned domains. In this sense, PubMed and Cochrane are specialised health sciences literature databases, whereas Scopus is a multidisciplinary database, where all knowledge fields are included. PubMed mainly contains journal and conferences articles, whereas Cochrane also contains reviews, trials, methods studies, and technology assessments, among others, with a total of 1,120,097 records. As a multidisciplinary database, we selected Scopus instead of Web of Science (WOS) because, besides being it easier to navigate, Scopus covers a superior number of journals than WOS, and most of the journals covered by WOS are also covered by Scopus [25]. Scopus covers more than 21,000 journal titles, more than 85,000 books, more than 6.5 million conference papers and 24 million patents [26].

Moreover, one advantage of PubMed and Cochrane is that we can perform the literature search using the Medical Subject Headings (MeSH) descriptors, whereas in Scopus we will use these descriptors as keywords. This manner the searches across all the databases will be homogenised to a great extent.

Table 3 summarises the size, temporal coverage, and domain for each of the selected electronic sources.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of records</th>
<th>Temporal coverage</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>24.6 million</td>
<td>1940 - present</td>
<td>Health Sciences</td>
</tr>
<tr>
<td>Cochrane</td>
<td>1,120,097</td>
<td>1992 - present</td>
<td>Health Sciences</td>
</tr>
<tr>
<td>Scopus</td>
<td>&gt; 30 million</td>
<td>1995 - present</td>
<td>Multidisciplinary</td>
</tr>
</tbody>
</table>

3. Results

3.1. Search results

Having defined our search strategy, the inclusion/exclusion criteria, the MeSH descriptors and information sources, we proceeded to search for the specific literature. We executed the same search in each of the electronic sources selected (i.e., PubMed, Cochrane, and Scopus), limiting to research works which their main topic deals with the MeSH descriptor (i.e., we set

\(^5\)http://www.ncbi.nlm.nih.gov/pubmed
\(^6\)http://www.cochranelibrary.com/
\(^7\)http://www.scopus.com/
the search with MeSH Major Topic). Table 4 shows the queries performed in the different databases.

Figure 1 summarises the flow followed during our search process and shows the number of articles obtained step by step after applying the different inclusion criteria. As it can be seen, the initial searches identified a pool of 1,583,134 total articles adding up the results of the three databases. We then applied the second and third inclusion criteria (Table 1), redefining the searches by performing the most relevant combinations of MeSH descriptors. Table 5 shows a summary of the final results for all the databases when all the inclusion criteria are applied.

Regarding PubMed database, we obtained 157 records when combining the descriptors Artificial Intelligence and Health Education together in the same query, thus forcing the publications to deal with these two topics. Moreover, restricting the query to take into consideration the third inclusion criteria, we found out that only 37 publications were identified with Natural Language Processing and Health Education, whereas only 1 publication explicitly used the descriptors Natural Language Processing and Health Literacy. Inspecting manually these records, we found out that the publication containing Natural Language Processing and Health Literacy was also included in the 37 records retrieved when health education was combined with Natural Language Processing, and these publications were at the same time, included in the 157 publications obtained for the more general query.

Focusing on the results obtained in Cochrane Library, we only obtained results when Artificial Intelligence was combined with Health Education. In this case, only 13 results were obtained comprising 1 review and 12 trials.

As far as Scopus results is concerned, we obtained 44 results when Artificial Intelligence was combined with Health Education; 15 when Natural Language Processing and Health Education were searched together; and only 6 when restricting the search to Natural Language Processing and Health Literacy. Since in this database we cannot perform the searches using the MeSH descriptors, there is no hierarchy associated to them, and therefore we will manually inspect all the retrieved articles.

So, we finally selected 235 distinct articles (157 from PubMed, 13 from Cochrane and 65 from Scopus) for further analysing them in detail.

3.2. Article analysis

Among the selected publications, we first carried out a preliminary scanning by reading the title and abstract, and based on it, we discarded 177 articles (120 from PubMed, 13 from Cochrane and 44 from Scopus), since none of them employed NLP techniques to address health literacy. Only a pilot study conducted in [27] manually analysed the impact of the inclusion health topics overview on reading comprehension, and measured the effect on a sample of 48 users. But, nevertheless, it did not use any NLP techniques or approach. From the remaining articles (58), we realised that it may occur that the same article could have been retrieved by several of the
<table>
<thead>
<tr>
<th>Database Search</th>
<th>Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE-ABS-KEY(&quot;Information Science&quot;) OR TITLE-ABS-KEY(&quot;Artificial Intelligence&quot;) OR TITLE-ABS-KEY(&quot;AI&quot;) OR TITLE-ABS-KEY(&quot;Machine Intelligence&quot;) OR TITLE-ABS-KEY(&quot;Natural Language Processing&quot;) OR TITLE-ABS-KEY(&quot;Health Education&quot;) OR TITLE-ABS-KEY(&quot;Consumer Health Information&quot;) OR TITLE-ABS-KEY(&quot;Health Literacy&quot;)</td>
<td>#3 or #1 or #2 or #6 or #7</td>
</tr>
<tr>
<td>all terms #1 to #7 MESH descriptor: Health Literacy explode all terms #8</td>
<td></td>
</tr>
<tr>
<td>Cochrane</td>
<td>#1 MESH descriptor: Information Science explode all terms #9</td>
</tr>
<tr>
<td>PubMed</td>
<td>#2 MESH descriptor: Artificial Intelligence explode all terms #10</td>
</tr>
<tr>
<td>Search</td>
<td>#3 MESH descriptor: Machine Intelligence explode all terms #11</td>
</tr>
<tr>
<td>Table 4: Initial search performed.</td>
<td>#4 MESH descriptor: Natural Language Processing explode all terms #12</td>
</tr>
<tr>
<td>OR TITLE-ABS-KEY(&quot;Health Education&quot;) OR TITLE-ABS-KEY(&quot;Consumer Health Information&quot;) OR TITLE-ABS-KEY(&quot;Health Literacy&quot;)</td>
<td>#5 MESH descriptor: Medicine MeSH explode all terms #13</td>
</tr>
<tr>
<td>OR TITLE-ABS-KEY(&quot;Artificial Intelligence&quot;) OR TITLE-ABS-KEY(&quot;AI&quot;) OR TITLE-ABS-KEY(&quot;Machine Intelligence&quot;) OR TITLE-ABS-KEY(&quot;Natural Language Processing&quot;) OR TITLE-ABS-KEY(&quot;Health Education&quot;) OR TITLE-ABS-KEY(&quot;Consumer Health Information&quot;) OR TITLE-ABS-KEY(&quot;Health Literacy&quot;)</td>
<td>#6 MESH descriptor: Information Science explode all terms #14</td>
</tr>
</tbody>
</table>
| OR TITLE-ABS-KEY("Information Science") OR TITLE-ABS-KEY("Artificial Intelligence") OR TITLE-ABS-KEY("AI") OR TITLE-ABS-KEY("Machine Intelligence") OR TITLE-ABS-KEY("Natural Language Processing") OR TITLE-ABS-KEY("Health Education") OR TITLE-ABS-KEY("Consumer Health Information") OR TITLE-ABS-KEY("Health Literacy")(OR TITLE-ABS-KEY("Information Science") OR TITLE-ABS-KEY("Artificial Intelligence") OR TITLE-ABS-KEY("AI") OR TITLE-ABS-KEY("Machine Intelligence") OR TITLE-ABS-KEY("Natural Language Processing") OR TITLE-ABS-KEY("Health Education") OR TITLE-ABS-KEY("Consumer Health Information") OR TITLE-ABS-KEY("Health Literacy")).))

Table 4: Initial search performed.
Figure 1: Information search flow diagram.
Table 5: Search performed and number of records retrieved applying all the inclusion criteria (as a query, we only provide the PubMed query, although for each database we generate the specific query with the appropriate syntax, but with the same semantics).

<table>
<thead>
<tr>
<th>Query</th>
<th>PubMed</th>
<th>Cochrane</th>
<th>Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>((((((“Artificial Intelligence”[MeSH Subheadings]) AND (“Health Education”[MeSH Subheadings]) AND ((English[Language]) OR (Spanish[Language]))))))</td>
<td>157</td>
<td>13</td>
<td>44</td>
</tr>
<tr>
<td>((((((“Natural Language Processing”[MeSH Subheadings]) AND (“Health Education”[MeSH Subheadings]) AND ((English[Language]) OR (Spanish[Language]))))))</td>
<td>37</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>((((((“Natural Language Processing”[MeSH Subheadings]) AND (“Health Literacy”[MeSH Subheadings]) AND ((English[Language]) OR (Spanish[Language]))))))</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

databases (at this stage from PubMed and Scopus). We reviewed this issue and this happened for 9 articles, so in the end, our final number of articles to analyse was 49.

Having a look at the temporal dimension through the publication year of the remaining 49 publications, we found out that the first attempts to apply NLP to health education date back to 1994. Between 2006 and 2008 the interest about these topics grows, although it was not until the last years (2013 and 2014) when the interest in these areas from a joint perspective mostly raised, as it can be seen in Figure 2.

![Figure 2: Publication date of the analysed publications.](image-url)
Further on, we carefully read and analysed the content of the 49 publications. It is worth noting that some of them, even though it seemed that by reading the abstract, NLP techniques for improving health literacy were involved, we realised that around 18% of the publications (9 of them) did not employ NLP for that purposes. They sometimes refer to NLP techniques, but indeed, they either provided an analysis and overview of the potentials and challenges of personalised medicine and the impact on NLP [28], or take advantage of Social Media for performing searches related to personalised health concepts [29]. In other cases, tools for including educational health materials, such as videos, pictograms, etc. that could help health literacy were developed [30], [31], but without using NLP.

For the remaining 82% of the publications, they were manually classified depending on its main goal and the purpose of using NLP techniques for health literacy. In this sense, four groups were created: i) development of NLP modules (question answering, information retrieval, information extraction, text summarization, text simplification, or natural language generation) that can help the better understanding of health information; ii) development of corpora, annotation schemes, vocabularies, non-expert terminology using NLP and health specific resources such as UMLS, or SNOMED-CT; iii) development of machine translation systems to help health literacy; and finally, iv) studies that analyse the readability of health information. Table 6 shows the percentage of publications and the specific references falling under each of these categories. This table also includes the publications that did not deal with NLP and health literacy at all (first row of the table).

3.2.1. Article analysis

Next, a more detailed analysis of the review publications according to the four groups established is provided.

- **NLP modules to help understand health information.** In this group we included the publications referring to NLP tasks (e.g., information retrieval, question answering, text summarization, text simplification, text generation, etc.) that were analysed either for helping users to obtain the information in which they were interested in, or knowing the behaviour of a user when interacting with health information.

Concerning information retrieval, most research works focus on the analysis and classification of how searches related to health information are performed by users [37]; how to improve the searches of specific medical information, such as vaccination-related information [39]; or analyse and classify the contents of Webpages containing health information [41], [45].

Regarding information extraction, the research work presented in [38] proposes a method to reliable extract different types of entities from
<table>
<thead>
<tr>
<th>Group</th>
<th>Percentage of Publications</th>
<th>Literature References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not dealing with NLP or health literacy</td>
<td>18.37%</td>
<td>[28], [29], [30], [31], [32], [33], [34], [35], [36]</td>
</tr>
<tr>
<td>i) NLP modules to help understand health information</td>
<td>51.02%</td>
<td>[37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61]</td>
</tr>
<tr>
<td>ii) Corpora, vocabulary and non-expert</td>
<td>14.28%</td>
<td>[62], [63], [64], [65], [66], [67], [68]</td>
</tr>
<tr>
<td>iii) Machine translation systems</td>
<td>4.08%</td>
<td>[69], [70]</td>
</tr>
<tr>
<td>iv) Readability analysis of health information</td>
<td>12.25%</td>
<td>[71], [72], [73], [74], [75], [76]</td>
</tr>
</tbody>
</table>

Table 6: Literature classification based on their main research goal and topic.
patient-authored text: symptoms and conditions, and drugs and treatments. They used lexico-syntactic analysis with seed dictionaries compiled from several Web resources (e.g. MedHelp\(^8\)) to identify these types of entities. In [42], a system that uses NLP to extract phrases, identify medical terms using the UMLS, and visualise the propositions, is developed. The purpose of this system is to reduce the amount of information a consumer must read by extracting concise information. In [55], a cross-lingual experiments (in French and Xhosa) in automatic detection of medical words that may be difficult to understand by patients is carried out, in order to detect similarity in both languages. NLP methods, such as classification techniques are used to reproduce manual annotation of the terms as understandable or not.

Other research approaches developing question answering systems that deal with health information can be found. This type of approaches mainly focus on helping users to find answers to their question related to medical information. For instance, in [40], AutoTutor tool is developed. This is a Web-based intelligent tutor system that interacts with users in natural language. The main goal of this tutoring systems is to engage users in tutorial dialogues to teach them about genetic breast cancer risk. Following the traditional pipeline in question answering systems, in [43], a bilingual French/English question answering system adapted to the health domain is developed. This system specifically focuses on the detection of the question's model, under the premise that this has a greater effect on the rest of the system. As it can be seen, this task is not addressed only for English, but also for other languages, such as Russian, Malayalam or Hindi. In [50], the evaluation of Russian question answering data in health domain is reported, where medical experts manually evaluated around 1,500 question answer pairs. Although the authors realised of the differences in the manual and automatic evaluation, they stress the usefulness of their proposed methodology for surveillance and health education campaigns. In [51], a question answering system that supports queries in Hindi and Malayalam, is developed. In this case, NLP techniques are employed for meaning extraction from the plain query and information from database is given back to the user in his/her native language. The authors claim that the proposed system can be effectively used in a wide range of application areas, including health. Also, the system proposed in [54] allows users to search for health information using natural language queries in a multilingual environment. For this, a basic health ontology is developed, and it is populated with health concepts in English, German and Turkish. Although the preliminary evaluation of the system was positively considered by users, its main limitation of this system is that the ontology has to be maintained by experts.

Regarding text simplification in [53], a tool to detect specialised medical terminology and provide explanations concerning is proposed. The

\(^8\)http://www.medhelp.org
evaluation of this tool was carried out with a group of users, and it was found out that this type of systems can improve the understanding of medical texts. Similar to the previous one, the research conducted in [56] outlines the problem with the language used in electronic health records, and develops the system NoteAid\textsuperscript{9}, that can automatically recognise medical concepts and link these concepts with consumer oriented, simplified definitions from external resources, such as MedlinePlus, UMLS, or Wikipedia. In this system, NLP is used mainly for sentence splitting, concept mapping, concept filtering and definition search. The results obtained showed that Wikipedia significantly improves the readability of the electronic health records, whereas the other external resources tested, MedlinePlus and UMLS can improve both content readability and content coverage for consumer health information. Moreover, in [57], a prototype for converting the language of electronic health records into a more accessible and plain language is designed and implemented. Its aim is to make the reports more comprehensible to consumers. Among the functionalities integrated in the prototype, one can found the identification of difficult terms, its replacement with easier synonyms, and the insertion of explanatory texts for them. These functionalities are developed thanks to the use of NLP tools, such as HITEx toolkit, as well as medical external resources, such as UMLS. Specifically, for concept extraction HITEx is used to parse the reports and map the terms to UMLS concepts. This step is necessary because they identify synonyms and related terms for a term based on the concept it represents. To avoid errors that might be introduced by disambiguation, the prototype does not deal with ambiguous terms. For synonym identification, the system finds the correspondence of UMLS concept to consumer health vocabulary concepts (CVH)\textsuperscript{10}, which are more understandable by non-expert users. Finally, for explanation generation, this is based on a familiarity score introduced by users, and depending on this score the application generates explanatory phrases based on the semantic relations in UMLS. The research work proposed in [60] analyses and develops a semi-automatic process, where difficult terms of a text can be simplified using NLP techniques and semantic ontologies such as Wordnet or UMLS. In particular, unfamiliar words are identified from texts based on their frequency, and a user evaluation was conducted for analysing the most appropriate substitution from several replacement candidates (e.g. hypernyms, definitions, or synonyms). However, this analysis and substitution was manually performed, since it is not trivial to automatically replace a word by a synonym due to possible genre and number agreement problems. Moreover, the effects of pronominal anaphora resolution were also manually analysed. Despite the manual analysis involved, this is one of the first research work that really employs NLP for improving health literacy. They

\textsuperscript{9}http://clinicalnotesaid.org
\textsuperscript{10}http://consumerhealthvocab.org/
showed how the difficulty of a text could be reduced by performing this type of simplifications, and how users perceived the text much easier to understand. This is only the first step in the process of semi-automatic text simplification, and there is still a lot of room for improvement. Furthermore, it is worth mentioning that these research works were carried out in the United States for the English language. This also indicates that similar approach need to be tested in other languages. Other type of simplifications involved the creation of tables of content [58] or the automatic extraction of paraphrases [59]. On the one hand, interactive tables of contents can help consumers access health information by providing an easy to understand structure, where NLP together with UMLS are employed for categorizing according to consumer-friendly labels for the UMLS semantic types and semantic groups. On the other hand, an automatic method for the acquisition of paraphrases for technical medical terms is proposed, under the hypothesis that paraphrases are easier to understand than original terms. The proposed approach is based on the morphological analysis of terms and on text mining of social media texts.

Finally, in this group we analysed different approaches addressing natural language generation in the medical domain. All the following approaches have a common goal, which is that of facilitating the communication between professionals and patients, and therefore different type of natural language generation systems are developed. As stated in [46], good communication is vital in health care, both among health care professionals, and between health care professionals and their patients. They also claim that the information may be easier to understand if it is described and explained in well-written documents rather than when only data is presented even in tabular or graphic form. On these basis, and under the premise that NLP techniques are matured enough to generate an automatic document from input data, they conduct an overview of different natural language generation techniques and systems that exist or could be applied to the health domain. Moreover, since there is a growing need for automated systems that can interview patients and consumers about their health and provide health education and behaviour change interventions using natural language dialog. In the last decades there have been developed different health dialog systems, and therefore, in [52] a survey of the theories, technologies and methodologies that are used in the construction and evaluation of these systems, along with a description of many of the systems developed and tested, is provided. Moreover, in [49] and [48], the issue of health literacy and more specifically, the difficulties in the communication between doctors and patients is also stressed. To mitigate this, the authors propose the use of advanced computer-based information systems to generate tailored, interactive handouts for doctors to communicate with patients. The goal of their approach is to generate natural language descriptions of migraine, its symptoms, triggering factors and prescriptions (MI-
GRAINE system). To achieve this they use text planning and user modelling techniques. Moreover, the systems is also capable of handling follow-up questions requesting further information, and generating responses in the context of previously supplied information. Similarly, the research work conducted in [47] tackles the generation of written explanations of a knowledge-based system in the domain of drug prescription in the framework of the OPADE project [77]. This system also integrates text planning techniques with surface generation. Another natural language generation system but in the field of reconstructive surgery can be found in [44]. In this case, the authors state that patient educational materials should be customized depending on the patient preferences. Therefore, they propose the development of a natural language generation system that can provide tailored information to patients in the abovementioned field. In particular, the authors proposed a tool that can guide surgeons in creating educational material, based on different textual variants extracted from a corpus, as well as the textual input the users enter in the system, which are stored in a database.

In this broad group, only one publication attempting to develop a recommender system for detecting relevant educational articles to patients was found out in [61]. The needs for a patient are extracted from his/her electronic medical records, and then topic modelling is used to identify and match topics. The MALLET and HITEX toolkits were used for determining the set of topics that describe a document, and to extract interesting terms from the document, respectively. The results obtained showed the difficulty of this task, since through the evaluation conducted, only for a few of them the recommendations received the maximum relevance rating.

- **Corpora, vocabulary and non-expert terminology resources development.**

Here, we grouped those articles that focus on the development of resources with the aim to increase the understanding of medical language and/or information.

UMLS is one of the most popular medical resources that is employed when developing NLP systems for the health domain. However, this is a very specialised resource, appropriate for health experts but not for non-expert professionals. In this sense, several approaches that try to map complex concepts to a more familiar terminology or develop terminologies adapted to patients can be proposed. In [66], NLP techniques are used to map propositions to UMLS terms with the aim to tailor personalised online information. In [68], an analysis of several strategies for the development of a consumer health vocabulary is conducted. Related to this idea, the main goal of [63] is to create a computer assisted update system that works with live corpora to identify new candidate terms for inclusion in consumer health vocabularies, in particular the standard open access and collaborative CHV
previously mentioned in the former group. Their proposed system was comprised of three modules: 1) a Web crawler and an HTML parser, 2) a candidate term filter that utilizes natural language processing tools including term recognition methods, and 3) a human review interface. The utility of the system was assessed by comparing the candidate term list it generated to a list of valid terms hand extracted from the text of the crawled webpages, showing that the proposed system was effective for generating a list of candidate terms for human review during CHV development. The development of annotated corpora is also very useful since it can help and improve the development of NLP tasks, such as natural language generation. In this respect, a coding scheme for a corpus of genetic patient letters was proposed in [67].

In other cases, the combination of graphical and textual information has been shown to be beneficial for helping in the understanding of complex terms. This is what is reported in [62], where using a combination of graphical pictures and text, the developed resource explains the salient concepts detected by NLP in the conclusion section of radiology reports. Also with the purpose of providing valuable information to non-expert health consumers, a visual analytic system, named DiseaseAtlas, is developed in [64]. This tool helps users navigate a large set of disease-related documents and understand multi-dimensional relationships for key semantic concepts such as symptoms and treatments. Real users evaluated this resource providing a very positive feedback and confirmed the main design objectives.

Finally, other type of Web 2.0 resources that have been also developed and which is very popular, is PatientsLikeMe [65], which is an online social networking community for patients. In this social network, community members can describe their symptoms to others in natural language terms, resulting in folksonomic tags available for clinical analysis and for browsing by other users to find patients with the same disease. The authors carried out an analysis of the use of this social network and found out that 43% PatientsLikeMe symptom terms were present as exact or synonymous terms in the UMLS metathesaurus; slightly more than half of the symptom terms either do not match the UMLS, or are unclassifiable; and SNOMED CT, accounts for 93% of the matching terms. The analysis of the failed matches revealed challenges for online patient communication, not only with healthcare professionals, but with other patients.

- **Machine translation systems.**

  In this group, we review systems that uses machine translation to facilitate the comprehension of medical information.

  Given the high-cost and impossibility of employing human translators, in [69], it is investigated whether multilingual machine translation could help make medical record content more comprehensible to patients who lack proficiency in the language of the records. In this approach, the analysis was limited to the use of Babel Fish, a popu-
lar general-purpose machine translation tool in order to translate 213 medical record sentences from English into Spanish, Chinese, Russian and Korean. Then, the comprehensibility and accuracy of the translation was evaluated, highlighting the fact that most of the translations were incomprehensible (76% to 92%) and/or incorrect (77% to 89%).

Having obtained these results, the authors also analysed the causes of the error, finally showing general-purpose machine translation tools like the Babel Fish may not be appropriate for the translation of medical records, and more sophisticated machine translation tools focus on this specific domain were needed.

Under the same motivation as the previous research work, in the study conducted in [70], a participatory design was used to model translation processes and inform the design of a public health translation tool. Specifically, the translation information workflow processes of two large health departments in Washington State were investigated. In this sense, researchers conducted interviews, performed a task analysis, and validated results with professionals to model translation workflow and identify functional requirements for the development of a prototype of a translation system.

- **Readability analysis of health information.**

  In this category we group those articles addressing the issue of health information readability, and the difficulty in understanding specialised medical information by non-expert users. The majority of the research works that fall into this group focus on the analysis of the readability for different types of medical information, analysing several linguistic metrics and stressing the differences compared to other types of information [71], [72], [74], [76], [75]. While it seems evident that there is a need to bridge the gap between specialised language and lay language, there are not many attempts to propose methods and techniques to bridge such gap. Only in [73], a method to build a comparable corpus of expert and non-expert medical French documents and to identify similar text segments of lay and specialised language is described. This is one of the first steps carried out in the literature that moves in this direction. The authors found that 59% of the text pairs were actually similar in both corpus and 37% were deemed exploitable for further processing, providing encouraging evidence for the target task of finding equivalent expressions between lay and specialised medical language.

4. Discussion

There is a common issue among all analysed publications: the need for and challenges of educating and informing patients are very well known, and these are even greater for patients with low levels of literacy, or health literacy. This has been mainly evidenced by the different readability analysis that have been conducted in the literature, and that have shown the existing gap and problems between the language employed by doctors and patients.
With the emergence of electronic medical records and computerised health information systems, there is an opportunity for automated tools to assist in addressing these challenges taking into account the maturity level reached by NLP tools, that can process natural language, which is the language employed for communicating between doctors and patients, even at a different level of specialisation. Doctors are familiarised with specific terminology which is often difficult to understand by patients, and therefore automatic or semi-automatic NLP processes can help on the task of facilitating such comprehension. As we found out from the revision carried out, great efforts have been done towards either the development of either NLP applications that aim to facilitate as much as possible the understanding of health information through text simplification, natural language generation, question answering or information extraction systems, among others, or the creation of consumer terminologies or vocabularies that are easier for non-expert users, and could help both health professionals and patients in their communication.

It is undoubtedly that although some research works have successfully addressed the fields of health literacy and NLP, there is still a lot of room for improvement, such as the one to be able to adapt the specialised information not only for users, but also taking into account the background and needs for specific users, and the type of disease, treatment, etc. This will imply that a new dimension has to be taken into account, which is that of personalised information in the sense of delivering to users only what they need and in the appropriate form and format. Nevertheless, this evidence-based review has served as an starting point for putting all the existing research concerning NLP and health literacy in context, and gaining some insights of the potentials challenges that still need to be faced in the coming years.

5. Conclusions

Health literacy is an important issue in our current digital information society, and there is an increasing interest in this research area, as it shows the extensive number of publications available in the literature. On the other hand, NLP has been applied to the medical and clinical domain in the recent years, showing great capabilities of the application of the developed technologies in these fields. However, it was not until 2013, when NLP applications started to focus on health literacy, developing modules and systems where advanced NLP techniques were applied to facilitate the understanding of medical documents to patients. Within such NLP techniques, we can find text simplification, an application which is gaining more and more importance since it aims at reducing the complexity in texts.

In light of this and motivated by the fact NLP can bring significant advantages to improving health literacy, we conducted a deep analysis and review of the research published concerning health literacy and natural language processing fields in an independent and a joint manner. Our main conclusion derived from the analysis carried out is that, whereas a wide range of approaches have been analysed and developed by each research area in an independent way, multi-disciplinary work considering such areas
together needs to be further exploited, being the advancements achieved in the area of text simplification or natural language generation a great starting point to be also applied in health literacy. More specifically, the high presence of publications concerning either health literacy or natural language processing, has been increasing during the last 5 years, thus indicating the relevance of these research areas.

As long as NLP techniques improve and resolve more tasks, they will become key techniques for being applied for health literacy, thus transforming complex terminology into a more accessible and understandable language that will lead to a better patient-doctor or patient-patient communication.

Authors’ contributions

Both authors equally contribute to the analysis and development of the research work carried out in this manuscript.

Conflicts of interest

Not applicable.

Acknowledgements

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, “Tratamiento inteligente de la información para la ayuda a la toma de decisiones” (GRE12-44), “Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario” (GRE13-15), DIIM2.0 (PROMETEOII/2014/001), ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), SAM (FP7-611312), and FIRST (FP7-287607).

References


[45] C. Weissenberger, S. Jonassen, J. Beranek-Chiu, M. Neumann, D. Müller, S. Bartelt,


