Building a Repository for Inferring the Meaning of Abbreviations Used in Clinical Studies

Efthymios Chondrogiannis*, Efstathios Karanastasis, Vassiliki Andronikou, Theodora Varvarigou
National Technical University of Athens, 9 Heroon Politechniou Str, 15773, Athens, Greece.

* Corresponding author. Tel.: 2107722132; email: chondrog@mail.ntua.gr
Manuscript submitted July 24, 2015; accepted December 30, 2015.
doi: 10.17706/jcp.12.1.76-88

Abstract: Abbreviations are widely used in clinical studies. Identifying their meaning is an essential step for further processing of the provided data. In this work, a novel system developed for detecting the meaning and context of the abbreviations specified in the whole corpus of clinicaltrials.gov studies is presented in detail. In the background, innovative algorithms and techniques were used for both acronym and non-acronym type abbreviation recognition purposes, while their evaluation was performed based on an abbreviations annotated corpus of clinical studies, which was purposely developed. The outcome of this work is a repository consisting of approximately 28 thousand abbreviations, while for each of them its possible senses along with their context have been recorded, so that it can be utilized for inferring the meaning of abbreviations met in existing and future documents in this domain. Further analysis of the recorded data indicated, amongst others, that there are considerable differences among abbreviations and their expansions or senses depending on the scientific domain in which they are used. Also, the abbreviations consisting of three and especially two characters are already highly ambiguous and their possible senses are going to be significantly increased in the following years.

Key words: Abbreviation recognition, clinical studies, sense repository.

1. Introduction

Abbreviations (ABRs) comprise an important part of a clinical study. They intend to provide short forms of often long texts (aka expansions) so that authors can efficiently use them in the rest of the document. However, in many cases, the ABRs are used without specifying their long form, which poses serious barriers to the further computer-based utilization of the provided data [1]. Identifying the meaning (aka sense) of an ABR presumes the detection of its possible senses and accordingly the selection of the appropriate one.

The majority of existing repositories provide the possible expansions (EXPs) or senses of an ABR, but they often do not contain a lot of evidences about their frequency of occurrence or the context in which each one is used, such as the ADAM database [2]. Hence, they cannot support tasks such as the process for resolving the meaning of the unspecified ABRs, especially when more than one frequently used EXP (i.e., used in a considerable amount of documents) exists for the same ABR. Nevertheless, there are a few publicly available repositories which adequately cover the context in which each sense is being used, such as the Acromine dictionary [3]. The latter has been developed based on the content retrieved from MEDLINE articles. As a result, the dictionary contains many unnecessary senses that are normally not met in clinical studies, while the overall number of times each sense has been used as well as its context may differ when used in clinical

studies. Moreover, since clinical trials is a "live" domain with approximately 20 thousand new studies being registered every year, new ABRs and/or senses are constantly being introduced, which may not be already covered by exiting repositories, especially if the latter are outdated.

Within this context, the design, construction and regularly update of an ABRs' repository based on the content retrieved from the clinical studies is necessary. In this work, we present a fully-automated system developed for detecting the meaning of the ABRs specified in clinicaltrials.gov studies [4]. In the background, innovative algorithms and techniques have been used for ABR recognition purposes, taking into account the "importance" of words, as highlighted in previous work of ours [5]. The system's evaluation has been performed based on an abbreviations-annotated corpus of documents that we developed. Subsequently, a thorough analysis of the collected data took place, taking into account the publication date of the documents, the length of the recorded ABRs and the techniques which are internally used for ABR construction/recognition purposes.

The document is structured as follows. In Section 2, the related work is presented. In Section 3, the overall approach followed for creating a repository based on the ABRs used in clinical studies is briefly described. In Section 4, the background mechanisms used for ABR recognition purposes are described. In Section 5, the evaluation of algorithms used is being presented, while, in Section 6, an analysis of the constructed ABR repository follows. Finally, in the last section the main points of this work are summarized and next steps are presented.

2. Related Work

For ABR recognition purposes, various algorithms and techniques have been proposed so far, which are classified in four categories, i.e. alignment-based, rule/pattern-based, machine learning-based and collocation-based [6]. A few representatives from each category are described in the next paragraphs.

Schwartz and Hearst [7] proposed a simple algorithm for detecting the EXP of an ABR based on the characters used. More precisely, in case the ABR is placed within parentheses the algorithm finds the shortest phrase that contains all of the ABR's characters (with the same order) on condition that the first character of both ABR and candidate-EXP match. Park and Bryd [8] used both text markers (e.g., parentheses and brackets) and cue words (e.g., acronym, short) for detecting candidate ABRs. Then a patterns-based approach was followed for detecting their EXP (if present) in the surrounding text, on condition that its length in words was lower than the value: min(|A| + 5, |A| * 2), where |A| is the number of the ABR's characters. Pustejovsky *et al.* [9] showed that the precision of an ABR recognition algorithm is significantly increased, if syntactic information is used to constrain the context in which to search for their EXPs.

Yu *et al.* [10] used a rule-based approach for detecting the full form of an ABR. More precisely, all possible Long Forms were created and accordingly the shorter one was selected that satisfied any of the specified matching-rules. Sohn *et al.* [11] used a variety of strategies for ABR identification purposes which were applied one after the other from the most reliable to the least reliable one. The most reliable strategy made an attempt to match an EXP with the ABR on condition that the later consisted of the first characters of the EXP's tokens. Several other strategies were also implemented taking into account the characters used along with stop words. Chang *et al.* [12] used a supervised machine learning approach for ABR recognition purposes. More precisely, they trained a binary logistic classifier based on a training set constructed from MEDLINE abstracts. In the background, 9 different features were used for calculating the alignment score, based on which the system decided whether the text provided was the EXP of the ABR or not. Kuo *et al.* [13] have also used a machine learning approach to identify abbreviations and definitions in biological literature, using a variety of features, including morphological, contextual, and numeric features as well as long form

composition.

Gawlik [14] has used a characteristic algorithm from each one out of the first three aforementioned categories against various datasets and has shown that despite the different approaches followed they all performed fairly well with an overall precision in the mid-90%, recall around 85%, and F-score near 90%. Ehrmann *et al.* [15] have also shown that existing state of the art algorithms can be used for ABR recognition purposes in 22 languages from different language families. The aforementioned techniques can directly provide the EXP of an ABR on condition that the ABR's characters match with the ones used in the EXP's tokens (aka acronym-type ABRs). However, in many cases, the ABR has partial or no similarity with its EXP for a variety of reasons. For instance, the ABR may come from another semantically equivalent phrase than the one provided (either in English or in any other language) or it may have been arbitrarily assigned, such as in the case of drug-codes.

For detecting the EXP in such cases (aka non-acronym-type ABRs), a statistics-based approach is necessary, such as the one presented by Zhou *et al.* [2]. In this approach, the text presented in the close vicinity of the ABRs is collected and accordingly processed for detecting those phrases which are used with a higher than specified frequency. Okazaki and Ananiadou [3] also collected the sentences with each ABR and accordingly used a modified "version" of a method called "C-value" which combines linguistic and statistical knowledge for calculating the likelihood of each candidate multi-word term being the EXP [16]. In general, statistics-based approaches demand a large amount of biomedical articles and especially a large number of re-definitions of the same ABRs for providing valuable results. Nevertheless, they can complement acronym-type ABRs recognition techniques by further processing the remaining "candidate" ABRs, as e.g. in the MDA system [17].

3. Overall Approach

For ABR-EXP detection purposes, approximately 190 thousand XML documents (DOCs) from the ClinicalTrials.gov site were downloaded and accordingly processed by the developed system, and especially those parts of a clinical study captured by free text (i.e., study title, description, eligibility criteria and summary). Throughout the aforementioned process, for each DOC the following data were also recorded: its unique ID (NCT number), the CT data provided date (first received date), and the medical conditions studied along with the interventions that took place (captured by Mesh Headings [18]). Fig. 1 presents the overall approach followed for detecting the meaning of ABRs used in the whole corpus of clinical studies.

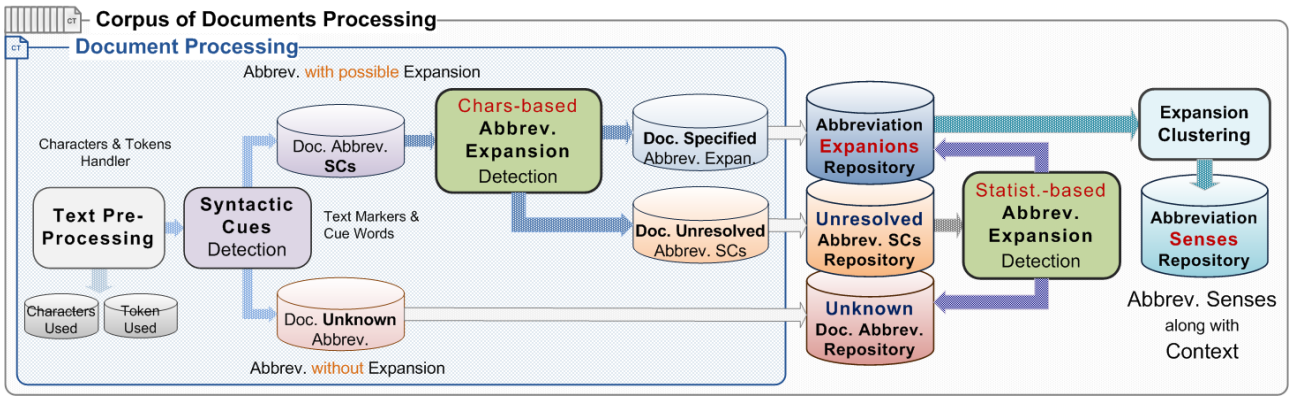

Fig. 1. The overall approach.

Initially, a pre-processing of the data specified in each clinical study is necessary in order to cope with the different symbols used for the same punctuation characters as well as possible tokenization issues. Accordingly, the text of each clinical study should be further examined for detecting the ABRs (e.g., ECG) or

generally the ABR's expressions (e.g., HIV-positive) used, especially the ones accompanied by their EXP. The latter, in the vast majority of cases, follow common patterns known as ABR syntactic cues (SCs) with the ones most widely used being those in which the ABR follows its EXP within parenthesis or brackets (SCs type A), or vice versa (SCs type B). The identified SCs were accordingly used for detecting the ABRs specified in the DOC. More precisely, for each one, it was examined whether the EXP had been provided in the text located exactly before or after the ABR or not, using a combination of alignment-based and rule-based techniques. Concerning the specified ABRs, since they often appear with the same sense in the DOC [19], the text preceding and/or following any occurrence of the ABR (in the same DOC) was also gathered.

The specified ABRs along with the text located in their close vicinity and the DOC that they belong to were stored in the ABR-EXP repository. The remaining SCs (i.e., the ones whose EXP could not be detected by using the aforementioned techniques), including the parts of the DOC in which the corresponding ABR was used, were placed in a separate repository named Unresolved ABR-SCs Repository. Finally, the ABRs used directly in a sentence that did not participate in any SC along with their surrounding text were stored in the Unknown DOC-ABRs Repository. The Unresolved ABR-SCs repository constructed from the analysis of the whole corpus of downloaded DOCs was further processed for detecting the EXP(s) of the ABRs, based on the co-occurrences of the same phrases. More precisely, for each ABR enclosed within parentheses or brackets, which often are used for providing the EXP of an ABR, the preceding text was collected and accordingly analyzed for EXP detection purposes. The detected pairs of ABR-EXP, along with the corresponding contextual data, were placed in the ABR-EXP Repository whereas the still unresolved ABRs were moved to the Unknown DOC-ABRs Repository.

For detecting the possible senses of an ABR along with the context in which each sense is used, the data residing in the ABR-EXP repository were further processed by placing the semantically equivalent EXPs in the same cluster based on their similarity, while the surrounding text along with the corresponding DOCs were used for detecting the context in which each sense is being used.

4. Background Mechanisms

4.1. Characters-Based Abbreviation Recognition

For detecting the EXP of the acronym-type ABRs, the SCs were carefully examined using (mainly) three different strategies (applied successively) which differ on the technique which was internally used for deciding whether the ABR matches with the candidate-EXP (i.e., the phrase the ABR may represent). In the first approach, it was examined whether all of the tokens participated in the construction of the ABR (Fig. 2(a)). In the second approach, it was examined whether the ABR matched with the EXP, if punctuation characters (e.g., commas) and especially one or more stop words [20] were ignored (Fig. 2(b)). Finally, in the third approach, it was examined whether an alignment could be achieved if one or more non-important words were additionally ignored (Fig. 2(c)).

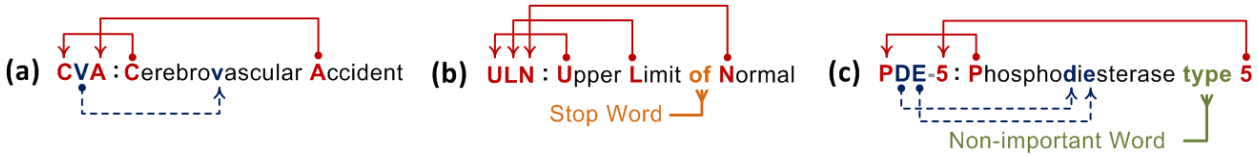

Fig. 2. Alignment among the ABR's and EXP's characters in three different examples.

The importance of tokens was measured taking into account the discrimination power of tokens (actually, the stem of words) used in the whole corpus of clinical studies. More precisely, the number of studies in which each token was used was counted and accordingly the token's importance was calculated as the

logarithm of the fraction: total number of studies divided by the number of studies the token is used in. Then, the values calculated were normalized so that they would belong in the range [0, 1]. Accordingly, a token was considered to be non-important, on condition that its value was lower than a predefined threshold. In fact, the specified threshold was closely associated with the ABR's length and varied between 0.2 (for 2-character-ABRs) and 0.4 (for rather long ABRs with more than 8 characters) [5].

In all of the implemented strategies, a slightly different process was followed depending on the SC used. More precisely, in case a SC-type-A was used, the shortest phrase that matches with the ABR provided was sought for, whereas, in case a SC-type-B was used, it was just examined whether the text enclosed within parentheses or brackets matched with the specified ABR, on condition that it did not start with the "e.g." or "i.e." Latin ABRs. In order to properly handle the data provided, it was taken into account whether the ABR was presented in singular or plural form (often denoted with the "s" character at its end) or whether it participated in an ABR expression (e.g., HIV-positive). In the first case, the "s" character was removed from the ABR, whereas, in the second case the corresponding word (i.e., "positive", in the aforementioned example) was eliminated from both the ABR and candidate-EXP (if present in the latter). The list of commonly used words emerged from a semi-automatic analysis of the compounds formed using ABRs, based on their frequency of appearance [5].

Concerning the technique that was internally used for matching purposes, the candidate EXP was initially processed for detecting the tokens which had contributed in the construction of the ABR. More precisely, it was examined whether the first character of the EXP's tokens appeared in the ABR in the same order (Fig. 2, solid lines). Then, it was examined whether the remaining ABR's characters, if any, were also present in the corresponding EXP's tokens in the same order (Fig. 2, dashed lines). In case that all of the EXP's tokens had contributed with one or more characters in the construction of the ABR, then the ABR was considered as "tightly" linked with the provided phrase (technique 1). Alternatively, on condition that all of the ABR's characters were aligned, the "non-aligned" tokens of the given phrase were further examined. In case they were stop words (technique 2) or generally non-important words (technique 3) the ABR was considered as "loosely" linked with the provided phrase.

In the aforementioned process, it should be noted that more than one "match" may have been feasible at the first step among the ABR and EXP characters. Consequently, all possible alignments were examined in order to find whether the ABR truly matched with the EXP. Also, numbers present in either ABR or EXP comprised a special case, according to which they should be considered the same independently of whether Roman or Arabic symbols had been used or even the corresponding words in English language. Moreover, in order to properly handle the letters that do not belong in the English alphabet, it was examined whether the ABR could be matched with the EXP if those letters were replaced with the corresponding English letters, internally specified (e.g., some Greek Letters mapped to the corresponding English Letters based on their sound or resemblance). For instance, using this approach the ABR "β-hCG" could be correctly matched with the EXP "beta-human chorionic gonadotropin".

In order to cope with some exceptional cases in which an ABR matches with the provided EXP if the order of the EXP's tokens is ignored, a fourth strategy was also implemented as follows. Firstly, all possible "phrases" were produced by ignoring the order of tokens that belonged to the provided candidate-EXP, and accordingly the aforementioned techniques were used for deciding whether they matched or not. It should be noted that the produced phrases were not allowed to start with a stop word (or punctuation character).

4.2. Statistics-Based Abbreviation Recognition

For detecting the EXP of non-acronym type ABRs, the text preceding the enclosed within parentheses or brackets ABR (i.e., SCs type A) was collected. More precisely, for each ABR a tree was initially constructed, based on the "unresolved" SCs that belong to the same ABR, which were accordingly used as a basis for

identifying their meaning. It should be noted that the SCs-type-B were only used when a limited number of SCs-type-A was available. The tree initially consisted of only one node, the root node (aka ABR node), which was accordingly updated based on the sequence of words existing before each ABR. The constructed tree indicates the number of times each word appears in the corresponding place before the ABR, on condition that the previous words are exactly the same.

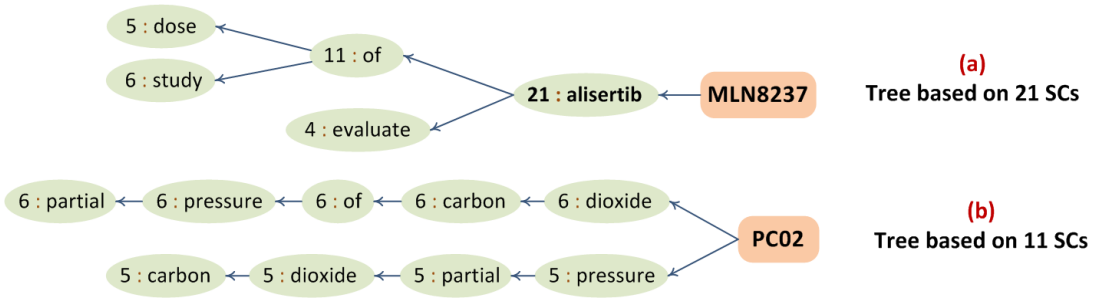

Fig. 3. The token-based trees constructed for (a) "MLN8237" and (b) "PC02".

Fig. 3(a) presents the automatically constructed tree for the "MLN8237" ABR, based on the text coming from 21 different SCs. The tree has been pruned for presentation purposes, by removing those edges and sub-graphs which were rarely used (e.g., used in less than 10% of the examples given). From this tree, it can be easily concluded that "Alisertib" is the EXP of the "MLN8237" ABR. However, in some cases, especially when the EXP consists of two or more tokens, there are many different ways to write the same phrase, which, in turn, hampers the ABR-EXP detection process using the constructed token-based tree (e.g., Fig. 3(b)), especially when there is a limited number of SCs for the corresponding ABR. The problem becomes much greater taking into account that in general, there is more than one EXP for the same sense.

In order to properly deal with the above examples, focus is given on the senses used. More precisely, starting from the leaves of the tree, the tree nodes are examined for detecting the semantically equivalent phrases (as described in the next section). The latter are the phrases formed from the nodes that should be traversed in order to reach to the top of the tree. In case the phrase was equal, the corresponding node was replaced by a "phrase" node which contains the semantically equal phrases, along with the number of times each phrase was used. It should be noted that the matching nodes along with their parents were removed, on condition that they were used exactly the same number of times; otherwise, remained in the tree but the corresponding numbers were replaced (i.e., reduced) with the updated ones.

For detecting the sense of the ABR along with the provided EXPs, the updated tree was further examined for finding the children of the ABRs node (including their sub-tree for rather long EXPs) which were used more times than a predefined threshold. Taking into account that the EXP of the ABR had already been detected, when an alignment was feasible, in this work the threshold was set to 75%, which practically means that only one sense could be detected for each ABR (i.e., the dominant one). However, the latter may be expressed with more than one different semantically equivalent EXP. Also, in order for the results to be valid, more than n SCs (in this work, n = 10, but this is part of the system configuration) should be provided for each ABR. Moreover, the EXP detected should contain at least one "important" word.

4.3. Senses and Context

For detecting the possible senses of each ABR, the recorded EXPs were examined. The clustering of the EXPs was based on a simple algorithm which iteratively examined the available ones for detecting those that were semantically equivalent. More precisely, an EXP was randomly selected and accordingly the remaining EXPs were processed for detecting the ones with the same meaning. The EXPs identified were placed in the

same cluster, while the whole process continued with the remaining EXPs (i.e., not already placed in a cluster). The clustering process terminated when all the EXPs were placed in semantically different clusters. For referencing purposes, the name of each cluster was the most widely used EXP (in singular form).

For detecting equal phrases, a plethora of techniques were used, such as elimination of punctuation characters and stop words. Also the Porter's stemming algorithm [21] was used for coping with terms variations (e.g., when the EXP was provided in plural form), while the importance of the remaining EXP's tokens was also taken into account, in order to properly handle the omission of some non-important words. For example, by taking into account the importance of tokens the EXPs "Upper Limit of Normal" and "Upper Limit of the Normal Range" could be correctly matched, since the importance of "range" is almost zero and hence it can be ignored. Concerning numbers, a purposely developed repository with all possible forms of a numeral was used, so that the corresponding tokens could be accurately matched (e.g., "two", "ii", "2", "second", "2nd"). Additionally, the ABRs which could possibly be present within an EXP were examined, as well as the possible tokenization mismatches among them. For instance, the EXPs "Continuous Intravenous Infusion" and "Continuous IV Infusion" of "CIVI" ABR are the same, since the ABR "IV" matches with the string "Intravenous". Finally, the Levenshtein distance metric [22] was used for detecting semantically equivalent tokens that either contain an error or generally have more than one accepted expressions with a very similar form from a characters point of view.

For each sense the DOCs were recorded in which the ABR was used with the corresponding meaning, which was accordingly utilised for calculating the number of times each sense was used across the whole corpus of clinical studies. For detecting the "broader" context of each sense, the Mesh Headings (i.e., Medical Conditions and Interventions) assigned to each DOC were collected, and accordingly a tree was created taking into account their hierarchy. More precisely, for each term, their parent nodes (based on the MeSH descriptors' hierarchy) were introduced in the constructed tree, also updating the number of times each node was present, in case it already existed in the tree. The text existing in the close vicinity of each sense (or ABR, on condition that it is being used with the same meaning) was further processed for detecting the different terms mentioned as well as the number of times each one appeared. In the current version, due to the enormous size of the collected text (i.e., more than 2 million sentence parts recorded) which requires a large amount of time and computer resources for detecting the unique concepts provided, focus was given on the distinct words used, and more precisely their stem, ignoring punctuation characters, numbers and stop words.

5. System Evaluation

In order to ensure the validity of the gathered data the implemented techniques were evaluated in advance based on an ABR-annotated corpus of clinical studies [23]. The latter was constructed based on a semi-automatic process of a few randomly selected DOCs from the clinicaltrialsregister.eu site. It should be noted that the aforementioned corpus of DOCs was used for evaluation purposes, since it was designed based on the content retrieved from clinical studies (141 randomly selected DOCs with more than 500 specified ABRs), while it also covers a variety of ABR-EXP matching cases, including a few examples in which the ABR has partial or non-similarity with the EXP provided from a characters point of view.

For evaluation purposes, the previously described system was used for detecting the EXP of the ABRs mentioned in each DOC and accordingly their outcome was compared with the ABR-EXP pairs specified by the user, calculating precision, recall, and F-measure (aka F-score). Precision is calculated by dividing the number of correctly detected ABR-EXP pairs by the total number of ABR-EXP pairs detected by the system. On the other hand, recall is calculated by dividing the number of correctly detected ABR-EXP pairs by the total number of ABR-EXP pairs specified by the user. F-measure is calculated as two times the product of

precision and recall divided by the sum of precision and recall.

The evaluation of the proposed system indicated that it can accurately detect the meaning of ABRs, with a precision of 0.9912, a recall of 0.8804 and an F-measure of 0.9325. Consequently, the automatically detected pairs of ABR-EXP were in almost every case correct, except from only one case where the system erroneously detected the EXP of the "hASCT" ABR. More precisely, the suggested EXP was "hematopoietic stem cell transplantation" (tight matches) whereas the one specified in the DOC was the "allogeneic hematopoietic stem cell transplantation" (which matches with the ABR if the tokens' order is ignored).

Concerning the missed ABR-EXP pairs (i.e., specified in the annotated corpus of DOCs but not identified by the algorithm), many of them had not been introduced in the DOC with one of the supported SCs. For instance, in some cases the ABR just followed the EXP (e.g., Alkaline phosphatase ALP > 2.5 x ULN). Also, in some cases, there was a distance between the ABR and EXP or additional information was provided within parentheses or brackets, which prevented the system from accurately detecting the "candidate" ABRs and accordingly their EXP. Moreover, a few pairs were missed due to partial matching among the ABR's and EXP's characters such as "HCV: Hepatitis C" (the word "virus" missing from EXP) and "5-FU: Fluorouracil" (the number does not appear in the EXP). Additionally, the corresponding ABRs were not mentioned frequently in the whole ABR-annotated corpus of DOCs and hence it was impossible to find their meaning using statistics-based approaches.

It should be noted that the precision and recall calculated refer to the whole system rather than each component separately. Consequently, they are highly affected by the outcome of the SCs detection process, while they are also affected by the two different ABR-EXP detection components used. In fact, the recall (and hence F-measure) of the system is slightly increased from 0.8605 to 0.8804 by also using the statistics-based approach. On the other hand, the precision of the system is not practically affected. Hence the F-measure of the system is increased from 0.9202 to 0.9325 respectively.

6. Results and Discussion

6.1. Abbreviations Repository

The constructed repository [24] contains approximately 28 thousand ABRs, while each one has 1.84 senses and is used in 15.18 different DOCs, on average. The ambiguity of the ABRs used in clinical studies is much lower than the one of biomedical ABRs. More precisely, the biomedical ABRs consisting of 2 up to 6 characters have 4.61 possible senses on average [12]. Consequently, the constructed repository simplifies the process of resolving the meaning of unspecified ABRs in clinical studies, since there is no need to cope with additional senses that in general are not met in clinical studies.

Table 1. Techniques Used for ABR-EXP Detection Purposes

ID	Description	Pairs	(%)
T1	All EXP's tokens have contributed with one (i.e., the first) or more (i.e., internal) characters in the construction of the ABR.	55540	72.26
T2	The ABR tightly matches with the EXP using technique T1 if one or more stop words are ignored.	10945	14.23
T3	The ABR tightly matches with the EXP using technique T2 if one or more non-important words are ignored.	7479	9.72
T4	The ABR tightly matches with the EXP using technique T3 if the order of EXP's tokens is ignored.	2389	3.11
ST	EXP detected using the statistics-based approach.	521	0.68

In Table 1 the techniques used internally for ABR-EXP detection purposes are summarized. As can be noticed, in a large amount of cases the ABRs are "tightly" linked with their EXPs (T1). In fact, the 58.36% of ABRs detected using T1 consist of the first characters of the EXP's tokens. Nevertheless, in a considerable amount of cases, the ABR is "loosely" linked with its EXP (T2 and T3). In such cases, the tokens are either

stop words or, in general, non-important words, in comparison with the other tokens of the EXP. For instance, in the ABR-EXP pair "CMV: Cytomegalovirus infection" the token “infection” is not so important since the meaning of the whole phrase can be inferred (or at least distinguished from the other senses of the ABR) based on the remaining EXP’s tokens. It should be further noted that stop words are often function words (e.g., auxiliary verbs, prepositions) which in turn are used for creating grammatically correct sentences or phrases. On the other hand, non-important words are content words such as nouns, verbs and adjectives, which help end users to form a picture in their mind about the topic under discussion.

Concerning the techniques used, some pairs could be also detected by ignoring the order of the EXP’s tokens, while in some cases it would also be necessary to ignore one or more stop or non-important words (T4). For example, “Corrected QT interval” matches with “QTc” if the non-important word “interval” is ignored along with the order of the remaining EXP’s tokens. Finally, in a limited number of cases, the EXPs were detected using a statistics-based approach (ST), such as “AZT: zidovudine” and “MK-5172: grazoprevir”. It should be noted that the ABRs resolved with ST are used in 40.88 DOCs on average, while in approximately 90% of SCs the provided sense was the one automatically detected.

6.2. Abbreviations with 2 or 3 Characters

The data residing in the ABR repository were further examined. In Table 2, the classification of ABRs based on their length (i.e., number of letters and digits) is presented. As can be noticed, the ABRs consisting of two characters are highly ambiguous with more than 8 different senses on average, while they are used in a large amount of DOCs. The ABRs with 3 characters have 3 different senses, while the ABRs with up to 3 characters are in general used with the same sense in the whole corpus of DOCs. Consequently, the meaning of ABRs with up to 3 characters can be easily detected, while the rest of ABRs (especially those with 2 characters) may be problematic, in terms of meaning inference.

Table 2. Classification of ABRs Based on Their Length, along with the Average (AVG) Number of Senses and DOCs Used

Characters Count	Abbreviations Count	Percent (%)	Senses (AVG)	Documents (AVG)
2	778	2.81	8.09	81.71
3	6738	24.31	2.98	33.17
4	10290	37.12	1.31	8.79
5	5129	18.49	1.17	6.04
6	2625	9.46	1.13	3.45
7	1241	4.48	1.12	2.37
8	508	1.83	1.11	2.93
9	281	1.01	1.05	4.31
10	137	0.49	1.04	2.92
ALL	27717	100	1.84	15.18

Further analysis of the ABRs consisting of two characters indicated that the 50.35% of their senses are used only in the specific DOC. Moreover, the 67.43% of senses are used in less than 5% of the DOCs in which the corresponding ABR is met. Hence, since the ⅔ of senses are only used locally, they can be practically ignored when searching for the meaning of the corresponding ABR, when the latter is not provided in the DOC. Also, the analysis indicated, that in a considerable number of cases, there is a dominant sense, which is used in almost every DOC. For instance, 1.03% of senses are used in 95% of the DOCs. Consequently, in the aforementioned cases it can be assumed that their meaning (if not provided in the document) is the dominant one, which is almost in every case the correct one. Nevertheless, there is a considerable number of ABRs with more than one widely used sense. Such cases are prone to errors, and the context in which they are used should be taken into account, especially when their meaning is not provided in the DOC.

The analysis of the context in which each sense is being used and especially of the tokens surrounding the aforementioned ABRs indicated that they are partially overlapping, and hence, they can be used for selecting the appropriate sense. For detecting the meaning of an ABR when its EXP is not provided in the document, a few approaches have been proposed [25], [26], which in general treat this problem as a Word Sense Disambiguation problem. This is a rather challenging topic, but it's out of the scope of this work.

6.3. New Abbreviations and Senses

Another important parameter examined is the introduction of new ABRs and senses in the previous 10 years. Initially, as part of this work, the clinical studies were organized based on the date their data were provided. Then, a temporal repository was created containing the ABRs and senses that were recorded based on the documents published before 2006. Accordingly, for each year after 2006, the newly introduced ABRs and senses were counted (Fig. 4) updating the temporal repository with the new ones, so that they are taken for granted for the next years.

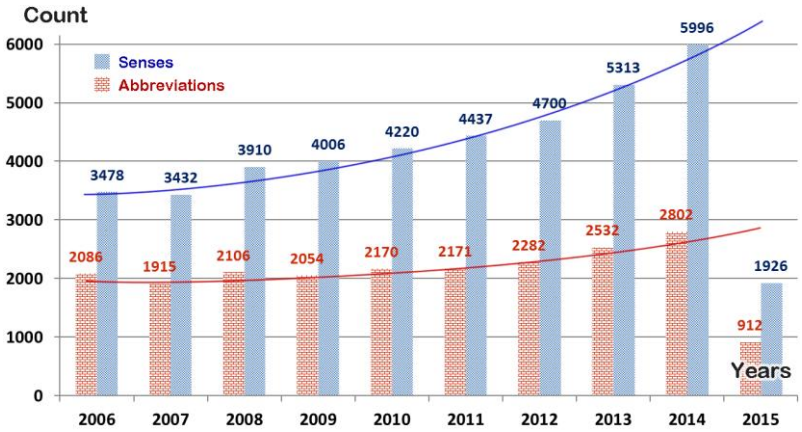

Fig. 4. The number of new ABRs and senses introduced every year.

As presented in Fig. 4, more than 2000 new ABRs and 3500 new senses were introduced in 2006, reaching 2800 and 6000 respectively in 2014, while the tendency of these numbers is to increase each year. Also, the number of senses is growing faster than the number of ABRs, since some of the new senses belong to existing ABRs. The above indicates that the constructed repository should be regularly updated based on the new clinical studies published. It should be noted that before 2006, only 6892 ABRs (i.e., 22.97% of total ABRs) and 9628 senses (i.e., 17.82% of total senses) had been recorded which is rational since approximately 25 thousands documents (i.e., 12.80% of total DOCs) were publicly available by then. Also, the number of new ABRs and senses for 2015 was approximately only the ⅓ in comparison to the respective number for the previous years, which is perfectly normal, since it is referring to DOCs that were downloaded by the end of March 2015.

Taking into account the fact that a) more than ½ of new senses (especially after 2010) belong to existing ABRs and b) the set of different ABRs that can be formed using a limited number of characters (mainly, letters and digits) is finite, it could be expected that their ambiguity is going to increase in the following years. Especially the possible senses of the 2-characters-ABRs are going to be increased dramatically, since 61.83% of possible 2-character-ABRs is already covered (in contrast for instance with 3-character-ABRs where only 15.74% is already covered). For this reason, care should be taken that the EXP of short ABRs are in general provided within the documents, the first time that they are used.

7. Conclusion and Next Steps

In this paper, a novel system was presented, developed for creating a repository containing ABRs specified

in clinical studies along with their meaning and the contexts in which their corresponding senses are used. In the background, innovative algorithms and techniques were implemented which could accurately detect the meaning of both acronym type and non-acronym type ABRs, also taking into account the importance of the tokens participating in their EXPs. The analysis of collected data indicated that ABRs have 1.84 different senses on average. However, ABRs with 3 and especially 2 characters are highly ambiguous, while their possible senses are going to be significantly increased in the following years.

As far as future work is concerned, the repository constructed is planned to be updated taking also into account specific words or phrases (i.e., linguistic cues) used for ABR's EXP specification purposes. Moreover, the sense detection process will be improved using a dictionary such as WordNet [27] as well as a more specific vocabulary such as UMLS [28]. Also, taking into account the distinctive role of the context in which the 2- or 3-characters-ABRs are used, the surrounding text will be further processed for detecting the unique concepts or patterns used with each one. Additionally, a system will be developed for detecting the most appropriate sense of an ABR when its EXP is not provided in the DOC.

Acknowledgment

This work is being supported by the OpenScienceLink project [29] and has been partially funded by the European Commission's CIP-PSP under contract number 325101. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

[1] Fred, H. L., & Cheng, T. O. (2003). Acronymesis: The exploding misuse of acronyms. *Texas Heart Institute Journal, 30(4)*, 255–257.

[2] Zhou, W., Torvik, V. I., & Smalheiser, N. R. (2006). ADAM: Another database of abbreviations in MEDLINE. *Bioinformatics, 22(22)*, 2813-2818.

[3] Okazaki, N., & Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics, 22(24)*, 3089-3095.

[4] ClinicalTrials.gov site. Retrieved May 1, 2015, from https://clinicaltrials.gov/

[5] Chondrogiannis, E., Andronikou, V., Karanastasis, E., & Varvarigou, T. (2015). Meaning inference of abbreviations appearing in clinical studies. *Proceedings of the 2015 Symposium on Languages, Applications and Technologies: Vol. 563* (pp. 127-136).

[6] Torii, M., Hu, Z., Song, M., Wu, C. H., & Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics, 8(Suppl. 9):S5*.

[7] Schwartz A., & Hearst, M. (2003). A Simple algorithm for identifying abbreviation definitions in biomedical text. *Proceedings of Pacific Symposium on Biocomputing* (pp. 451-462).

[8] Park, Y., & Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 126–133).

[9] Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., & Morrell, M. (1001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Stud Health Technol Inform., 84*, 371-375.

[10] Yu, H., Hripcsak, G., & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association, 9(3),* 262–272.

[11] Sohn, S., Comeau, D. C., Kim, W., & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics, 9*, 402.

[12] Chang, J. T., Schutze, H., & Altman, R. B. (2002). Creating an online dictionary of abbreviations from

MEDLINE. *Journal of the American Medical Informatics Association, 9(6),* 612–620.
[13] Kuo, C., Ling, M., Lin, K., & Hsu, C. (2009). BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics*.
[14] Gawlik, M. (2010). Comparison of abbreviation recognition algorithms.
[15] Ehrmann, M., Rocca, L., Steinberger, R., & Tanev, H. (2013). Acronym recognition and processing in 22 languages. *Proceedings of the 9th Conference Recent Advances in Natural Language Processing* (pp. 237-244).
[16] Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing, 6*, 145–179.
[17] Xu, Y., Wang, Z., Lei, Y., Zhao, Y., & Xue, Y. (2009). MBA: A literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics, 10,* 14.
[18] Medical Subject Headings (MeSH) site. Retrieved May 1, 2015, from https://www.nlm.nih.gov/mesh/
[19] Gale, W. A., Church, K. W., & Yarowsky, D. (1992). *One sense per discourse. Proceedings of the Workshop on Speech and Natural Language HLT '91* (pp. 233-237).
[20] Common English Words. Retrieved May 1, 2015, from http://www.textfixer.com/
[21] Porter, M. F. (2006). An algorithm for suffix stripping. *Program, 40(3),* 211-218.
[22] Pieterse, V., & Black, E. P. (1999). Levenshtein distance. *Dictionary of Algorithms and Data Structures*, CRC Press LLC.
[23] An Abbreviations-annotated Corpus of Clinical Studies. Retrieved May 1, 2015, from http://ponte.grid.ece.ntua.gr:8080/AbbrAnnotatedCorpus/
[24] A Repository with Abbreviations Specified in ClinicalTrials.gov studies. Retrieved May 1, 2015, from http://ponte.grid.ece.ntua.gr:8080/ClinicStudiesAbbrev/
[25] Stevenson, M., Guo, Y., Abdulaziz, A., & Gaizauskas, R. (2009). Disambiguation of biomedical abbreviations. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 71-79).
[26] Xu, H., Fan, J. W., Hripcsak, G., Mendonca, E. A., Markatou, M., & Friedman, C. (2007). Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics, 23(8),* 1015-1022.
[27] Miller, A. G. (1995). WordNet: A lexical database for English. *Commun. ACM, 38(11),* 39-41.
[28] Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research, 32,* D267–D270.
[29] OpenSicenceLink EU project. Retrieved May 1, 2015, from http://opensciencelink.eu/

Efthymios Chondrogiannis received his diploma from the Department of Electrical and Computer Engineering of the National Technical University of Athens (NTUA), Greece, in 2008. He completed his thesis with the title "Analysis and Integration of the Operating System Android in Grid" to the sector of Telecommunication, Electronics and Computer Systems of the NTUA. He is currently pursuing his PhD in grid computing at the Telecommunications Laboratory of the Department of Electrical and Computer Engineering of the NTUA and he is a researcher for the Institute of Communication and Computer Systems (ICCS). In the past, he has completed several IT and Web Projects in the private sector. Also, he has participated in a few European Projects such as OpenScienceLink and PONTE. His research interests include information engineering, distributed systems, service oriented architectures, bioinformatics, health information systems, semantic interoperability, healthcare and clinical research standards, ontologies, and text mining.

Efstathios Karanastasis received his diploma in electrical and computer engineering from the University of Patras, Greece, in 2007. In the past has undertaken IT and web projects in the private sector. In 2004 he worked for Athens Olympic Broadcasting, in the production and archiving of the broadcasted program of the XXVIII Summer Olympic Games. During the course of his military service he worked for the Center of Informatics of the Greek Army (KEPYES) as a software developer. In addition, he qualified as a consultant IT-specialist at the e-government team assembled by the Greek Ministry of Administrative Reform and e-Governance in cooperation with the Ministry of Defense. Currently, he is a PhD candidate in the Department of Electrical and Computer Engineering of the National Technical University of Athens (NTUA) and has been employed since 2006 as a researcher at the Institute of Communications and Computer Systems (ICCS). He has participated in numerous EU-funded IT projects including HPC-Europa, AKOGRIMO, BEinGRID, PONTE and OpenScienceLink, mainly involved with SOA platforms design and implementation, as well as grid and high performance computing. He is fluent in Greek, English and German. His research interests include service oriented architectures, knowledge modeling, data integration, grid and cloud computing, and web portals.

Vassiliki Andronikou received her diploma from the Electrical and Computer Engineering School of the National Technical University of Athens in 2004. She has worked in the National Bank of Greece and the Organization of Telecommunications of Greece, while since 2004 she has been a research associate and PhD candidate in the Telecommunications Laboratory of the NTUA. In 2005 she was given the Ericsson Award for her thesis on "Mobile IPv6 with Fast Handovers". In 2009, she received her PhD in the area of biometric systems focusing on innovative techniques for the improvement of their efficacy and effectiveness at fusion and resources level from the School of Electrical and Computer Engineers of NTUA. Her research has involved her participation in many European projects, such as OpenScienceLink, PONTE, BEinGRID, POLYMNIA, FIDIS and AKOGRIMO, with her interests focusing on the fields of fusion in multimodal biometric systems, knowledge modeling in clinical research and EHR interoperability.

Theodora A. Varvarigou received the B.Tech degree from the National Technical University of Athens, Athens, Greece in 1988, the MS degrees in electrical engineering (1989) and in computer science (1991) from Stanford University, Stanford, California in 1989 and the Ph.D. degree from Stanford University as well in 1991. She worked at AT&T Bell Labs, Holmdel, New Jersey between 1991 and 1995. Between 1995 and 1997 she worked as an Assistant Professor at the Technical University of Crete, Chania, Greece. Since 1997 she was elected as an Assistant Professor while since 2007 she is a Professor at the National Technical University of Athens. Prof. Varvarigou has great experience in the area of embedded systems and cloud computing. She has published more than 150 papers in leading journals and conferences. She has participated and coordinated several EU funded projects such as PONTE, OpenScienceLink, ALLADIN, COSMOS, SocIoS, CONSENSUS and SCOVIS.s.