A Web Corpus for eCare: Collection, Lay Annotation and Learning -First Results-

Marina Santini*, Arne Jönsson†*, Mikael Nyström†*
*RISE SICS East, †Linköping University
Linköping, Sweden
marina.santini@ri.se, arne.jonsson@liu.se, mikael.nystrom@liu.se

Marjan Alirezai
Örebro University
Örebro, Sweden
marjan.alirezai@oru.se

***Abstract*—In this position paper, we put forward two claims: 1) it is possible to design a dynamic and extensible corpus without running the risk of getting into scalability problems; 2) it is possible to devise noise-resistant Language Technology applications without affecting performance. To support our claims, we describe the design, construction and limitations of a very specialized medical web corpus, called eCare_Sv_01, and we present two experiments on lay-specialized text classification. eCare_Sv_01 is a small corpus of web documents written in Swedish. The corpus contains documents about chronic diseases. The sublanguage used in each document has been labelled as "lay" or "specialized" by a lay annotator. The corpus is designed as a flexible text resource, where additional medical documents will be appended over time. Experiments show that the lay-specialized labels assigned by the lay annotator are reliably learned by standard classifiers. More specifically, Experiment 1 shows that scalability is not an issue when increasing the size of the datasets to be learned from 156 up to 801 documents. Experiment 2 shows that lay-specialized labels can be learned regardless of the large amount of disturbing factors, such as machine translated documents or low-quality texts, which are numerous in the corpus.**

I. Introduction

BUILDING a very specialized medical corpus bootstrapped from the web is not a trivial task. The web is indeed a rich textual resource, but it contains many irrelevant or low-quality documents (noisy texts) that may affect the overall usefulness of a corpus. Sorting out noisy documents from the good ones is a tedious, expensive and time-consuming task. Additionally, in the medical field there is often the need to update a document collection with the latest illness-related texts, containing novel findings, new issues or unprecedented cases.

Web corpora are often at the core of Language Technology applications (henceforth LT applications). Since the design and the quality of web corpora affect the reliability and the performance of corpus-based LT applications, we investigate alternative approaches to traditional corpus design and propose an approach that can ensure robustness without affecting the overall performance. In particular, we focus on the relations between performance, scalability and noise on a specific LT task, namely lay-specialized text classification.

Performance can be affected by both scalability issues and by noise. We use the word "performance" to refer to the results achieved on a specific task (e.g. text classification), while with the word "scalability" we refer to the application's ability to adapt to the growing size of the underlying corpus without requiring major design changes. Scalability and performance are often associated, because performance can be affected by scalability issues.

Noise, on the other hand, is pervasive in Language Technology. Normally, LT applications are developed to handle clean texts. These applications may suffer from a significant performance decline when increasing the noise level of the corpus. Cleaning texts or removing noisy documents from a corpus is often a daunting and expensive task. With the expressions "noise resilience" and "noise-resistant LT applications", we refer to the property of keeping up a good performance in the presence of noisy documents.

In this position paper, we argue that: 1) designing dynamic and extensible web corpora does not necessarily imply scalability issues for LT applications; 2) including noisy texts in a corpus does not necessarily imply decreases in performance. Robustness to scalability issues and to noise are desirable qualities of any LT applications.

To support our claims, we describe the design, the construction and the limitations of a very specialized medical web corpus, called *eCare_Sv_01*, and we explore the case of lay-specialized text classification. *eCare_Sv_01* is a small corpus of web documents written in Swedish. The documents in the corpus contain specialized medical terms. The sublanguage used in each document has been labelled as *lay* or *specialized* by a lay annotator. The corpus is designed as a flexible text resource, where additional medical documents will be appended over time.

The creation of *eCare_Sv_01* stems from the following needs: (1) having a publicly-available medical corpus annotated with lay-specialized labels that can be easily shared; (2) having a corpus with a design and a structure that allow for expansion with additional documents over time; (3) accounting for very specific medical terms.

To date (see Section IV), going to specific websites and dumping lay-specialized medical texts is what is being normally done. As a matter of fact, these websites do not contain all the illnesses but only the most common ones. The same is true for user-generated texts, such as those that can be found in forums and blogs, since users mostly talk about general problems or common diseases. Another common approach has been to focus on journals or, more rarely, on patient record collections, but in this case there exist copyright, ethical and legal restrictions that limit the shareability and experimental replicability. For all these reasons, with *eCare_Sv_01* we are exploring a different avenue. More specifically, with *eCare_Sv_01* the idea is to pre-select some very specific medical terms (not just the most common illnesses), use them as seeds in a search engine and download only the pages that are related to the specific terms we focus on. In practice, we aim at building a corpus that contains documents that are related **only to specific medical terms** that indicate chronic diseases, and that are not always documented in medical websites, such as the Swedish medical information portal called "1177 Vårdguiden".

As mentioned above, in this paper we describe only preparation work, i.e. the construction of the corpus and present first results. However, the long-term work that leverages on *eCare_Sv_01* include tasks such as: (1) the definition and measurement of the domain-specificity of a corpus (that we call "domainhood"); (2) the automatic extraction of lay-specialized medical lexica and the creation of lexical ontologies from texts that contain specialized terms and lay synonyms; (3) the development of machine-learning-based medical LT applications (e.g. multi-labelled, semi-supervised, weakly-supervised and unsupervised lay-specialized text classification).

The current version of *eCare_Sv_01* contains Swedish web documents related to chronic diseases that are classified as such in the SNOMED CT ontology[1]. Since chronic diseases can be treated at home and monitored through electronic devices (sensors, self-reported records, etc.), we intend to use the corpus for LT experiments within the *E-care@home* project (see Section II). In future, the corpus will be expanded with additional diseases (e.g. "tachycardia" or "dementia"), not necessarily classified as chronic in SNOMED CT.

II. Lay-Specialized Medical Terminology and the Internet of Things

E-care@home[2] is a multi-disciplinary project investigating how to ensure medical care at home and avoid long-term hospitalization in the eldercare. Long hospitalizations are discomforting for elderly patients and expensive for the national healthcare system. Providing medical care at home to the elderly can be effective by populating the home with electronic devices (“things”), i.e. sensors and actuators, and linking them to the Internet. Creating such an Internet-of-Things (IoT) infrastructure is feasible and profitable [1]. Information gathered by sensors are lists of numbers. It is possible however to convert these bare numbers into specialized semantic concepts [2]. This conversion complies to one of *E-care@home* major objectives, i.e. to represent information in a "human consumable way". Converting numbers into concepts expressed in a natural language that experts can understand is certainly a big step forward and it is especially valuable for health professionals, who can use this converted information for timely decision-making. Additionally, since in the *E-care@home* framework patients are empowered and take active part in the management of their illnesses, it is no longer enough to convert sensor data to a medical language that only experts understand. Patients too should be included in the information cycle. There are linguistic obstacles, though. As a matter of fact, medicine is a domain where there exists a divide between the language used by health professionals and the language normally used and understood by patients, caregivers or relatives. This is a well-known problem that is extensively researched (see Section IV). In the project, it is pointed out that: "Patients and citizens will be faced with the technical language of the professional health records. Health care professionals are faced with issues of trustworthiness of personal health record data." Here lies the motivation of *eCare_Sv_01*: the construction of *eCare_Sv_01* exemplifies how to build a concept-specific medical corpus that is useful for eHealth and eCare-oriented LT applications, such as the automatic extraction of lay synonyms corresponding to medical terms.

III. Lay vs Specialized Sublanguage

The need of lay synonyms or lay paraphrases that match specialized medical terminology used by healthcare professionals has been the focus of recent research, both in Language Technology [3], and in the clinical community [4]. Research on lay-specialized sublanguages is brought about by the need to improve communication between two specific user groups: the layman on one side, and the expert on the other side. A classical example of a medical term is "varicella", which patients often call "chickenpox". The word "varicella" is a specialized medical term, while "chickenpox" is a lay synonym.

To date, there is no agreed lexical expression that subsumes concepts such as "lay", "normal", "simplified", "expert", "specialized", "consumer health vocabulary", "consumer terminology", and the like. Researchers use different expressions to indicate these kinds of language varieties, for instance, "different genres (such as specialized and lay texts)" [5]; "discourse types (lay and specialized)" [3]; or "registers" [6]. Most commonly, however, researchers do not relate the specialized-lay varieties to any superordinate category, as in [7].

Instead of using an umbrella term like "discourse", or employing textual dimensions like "register" or "genre", we suggest adopting the category **sublanguage** to refer to the different language varieties employed by user groups in certain situational or communicative contexts. Normally, a sublanguage refers to a specialized language or jargon associated with a specific user group, (e.g. the jargon used by teenagers stored in the Corpus of London Teenagers [8]) or to a very specialized domain-specific communication style (e.g. the "notices to skippers"). Computationally, a sublanguage is charac-

[1] International SNOMED CT Browser: http://browser.ihtsdotools.org/

[2] Project website: http://ecareathome.se/

terized by domain-specific terms (or word co-occurrences) and syntactic cues that deviate from normal language use [9]. *We broaden this definition of sublanguage in order to encompass language varieties that are commonly used when two or more user groups communicate in specific domains or in special communicative contexts*. Arguably, this definition of sublanguage is unambiguous and applicable to all the domains where the domain-specificity of a jargon causes some kind of "diglossia", and a gap in human communication. Following the extended definition, we can then say that in the medical domain, two sublanguages normally come in contact, namely the **lay sublanguage** used by patients and their relatives (the lay) and the **specialized sublanguage** used by healthcare professionals (the expert).

Normally, lay synonyms are based on everyday language, and are easier to read and to understand than medical terminology, which conversely have high-brow connotation. For normal people without a medical education or background, medical terms are often opaque or hard to remember due the Greek and/or Latin etymology. These terms are called "neo-classical" terms, and, interestingly, recent research shows that also healthcare professionals tend to "normalize" this type of lexicon to everyday language, as in the case of "Swedification" of Latin and Greek affixes in patient records [10]. Generally speaking, it seems that the "layfication" of medical language is an extensive phenomenon that affects, in different ways, several user groups.

IV. Previous Work

The automatic identification and extraction of specialized medical terminology and its systematization and standardization is an ongoing effort in many languages. For the Swedish language, experiments show that semi-automatic methods are reliable and can be implemented in real-life settings [11], [12].

Since the focus of our research is on the lay sublanguage rather than on the systematization and standardization of expert terminology, in this section we focus on the latest research on how lay-specialized medical text collections have been designed or used in several languages.

Examples for the English language include a method to mine a lexicon of medical terms and lay equivalents using abstracts of clinical studies and corresponding news stories written for a lay audience [13]. The collection is structured as a parallel corpus of documents for clinicians and for consumers. The study presented in [14] focuses on the linguistic habits of consumers. In this study, the authors empirically evaluate the applicability of their approach using a large data sample consisting of MedLine abstracts as well as posts from a popular online health portal, the MedHelp forum. The "propensity of a term", which is a measure based on the ratio of frequency of occurrence, was used to differentiate consumer terms from professional terms.

For French, experiments have been carried out by [3] to build lay-specialized monolingual comparable corpora using web documents belonging to specific genres from public websites in the medical domain. The corpus devised by [3] is quite different from *eCare_Sv_01*, since [3] include in their corpus various texts containing any kind of medical terminology, while in the design of *eCare_Sv_01* we only focus on texts related to *very specialized illnesses*, i.e. those listed under the chronic diseases node in the Swedish SNOMED CT.

In Sweden, research on medical collections is well-established and thriving. For instance, [6] created a unique medical test collection for Information Retrieval to provide the possibility to assess the document relevance to a query according to two user groups, namely patients or doctors. The focus of [7] is on the simplification of one single genre, namely the medical journal genre. To this purpose, the authors used a subset of a collection built from the journal Svenska Läkartidningen, i.e. the Journal of the Swedish Medical association, that was created by [15]. Another unique language resource is the Stockholm EPR (Electronic Patient Records) Corpus [16], [17], which comprises real data from more than two million patient records.

The medical text collections briefly described above are important language resources that, although not always publicly available, can be shared for research purposes under certain conditions. With *eCare_Sv_01*, we are exploring an alternative research path, where a text corpus is purposely designed to be publicly available.

V. The Corpus

eCare_Sv_01 is a small text collection bootstrapped from the web. It contains 801 web documents that have been labelled by a *lay annotator*. In the following subsections, we describe its construction and the actual corpus.

A. Seeds

We started off with approximately 1300 term seeds designating chronic diseases in the Swedish SNOMED CT. A qualitative linguistic analysis of the term seeds revealed a wide range of variation as for number of words and syntactic complexity. For instance, multiword terms (n-grams) are much more frequent than single-word terms (unigrams). We counted 13 unigrams (see Table I), 215 bigrams, and the rest of the seeds were characterized by specialized terms and complex syntax, such as: "kronisk inkomplett tetraplegi orsakad av ryggmärgsskada mellan femte och sjunde halskotan" (English: "Chronic incomplete quadriplegia due to spinal cord lesion between fifth and seventh cervical vertebra").

To bootstrap the corpus we used unigrams and bigrams only. This decision was based on the assumptions that (1) unigram- and bigram-terms are more findable on the web than syntactically complex keyword seeds, and (2) complex multiword terms are less likely to have a lay synonym or paraphrase. It should be noticed however that Swedish is a compound language where several words are united in one single graphical unit, so the distinction between unigrams and bigrams is sometimes blurred.

B. Preprocessing and Download

A preliminary investigation showed that when searching for medical terms (the seeds) as search keywords, the list of results

TABLE I
UNIGRAM SEEDS

Seeds (Swedish)	Translation (English)	SCTID
ansiktstics	Facial tic disorder	230335009
bukangina	Abdominal angina	241154007
chalcosis	Chalcosis	46623005
fluoros	Fluorosis	244183009
kromoblastomykos	Chromoblastomycosis	187079000
lipoidnefros	Minimal change disease	44785005
lungemfysem	Pulmonary emphysema	87433001
mycetom	Mycetoma	410039003
ozena	Ozena	69646003
polyserosit	Polyserositis	123598000
postkardiotomisyndrom	Postcardiotomy syndrome	78643003
Swimmingpooldermatit	Swimming pool dermatitis	277784005
trumhinneatelektas	Tympanic atelectasis	232258001

contains many irrelevant documents, which make a specialized corpus noisy. We decided to use seeds in the following way. Each seed was used as search keyword in *Google.se* (Google web domain for Sweden). For each seed, Google returned a number of hits. We limited our analysis to hits on the first page. We manually opened each snippet to have an idea of the type of web documents that were retrieved. For each search lap, several documents were irrelevant and several were duplicated. 74 keyword seeds were discarded because the retrieved documents were irrelevant or written in a foreign language. Unsurprisingly, we also noticed that the number of retrieved pages depends on how common a disease is. For instance, "ansiktstics" (English: “facial tics”) had many hits, while "chalcosis" (English: “chalcosis”) very few. As a rule of thumb, we decided to select a maximum of 20 documents for the most common illnesses, and as many as we could for rarer diseases. After this preprocessing phase, we applied BootCat [18] using the advanced settings (i.e. *url seeds*) to create the web corpus.

We handed out the bootcat-ted documents to a native Swedish speaker (an academic) who does not work in the medical domain and has no medical-related education. The lay annotator proceeded with the labelling by applying a *lay* or *specialized* label to each text in the corpus.

C. eCare_Sv_01 in a Nutshell

eCare_Sv_01 has been bootstrapped using 228 terms (13 unigrams and 215 bigrams). After the preprocessing, 843 urls (112 for unigrams and 731 for bigrams) were factored out and used as *url-seeds* in BootCat. Some of the urls were automatically discarded by BootCat (e.g. bilingual documents were discarded) and some bootcat-ted documents were empty. Finally, 801 documents were successfully bootcat-ted. Table II shows the corpus statistics.

The annotator pointed out that the writing quality of a number of web documents was poor, mainly because they had been machine translated, and not written by humans. Some of the web documents explicitly stated "Översatt från engelska av Microsoft" (English: Translated from English by Microsoft). Out of 801 web documents, 339 have received comments by the lay annotator, e.g. "Machine Translated" or "it is about animals and not about humans".

Essentially, we can observe that the corpus is noisy. The annotator’s comments help us understand the different types of noise and emphasize a crucial issue that is underexplored in corpus- and computational linguistics, i.e. the reliability and the quality of corpora bootstrapped from the web. The automatic discrimination of "good" documents from "bad" ones is an important problem, especially in sensitive domains like the medical or legal domains. This topic will certainly be explored in our future research. However, in the experiments that we report in this paper we took another perspective and investigated to what extent lay-specialized text classification is robust to noise. Since cleaning a corpus might be prohibitive for a number of reasons, the challenge is to see whether noisy corpora can be used in Language Technology without affecting the performance of LT applications.

For this reasons, the noisy documents have been left in the corpus but they are flagged so they can be easily included or bypassed, according to the purpose of the research, as we did in the experiments presented in Section VII. Other types of research that can benefit from the inclusion of noisy texts in the corpus include the automatic analysis of MT "translationese" [19] and the automatic quality assessment of text writing[3].

D. Web Corpora and Copyright

Legislation about the copyright and the re-usability of web documents that are not licensed under a Creative Commons Attribution has not been standardized globally. In some countries, regulations exist, but they are supposed to be valid on the national territory. For instance, in the US, the legal doctrine called Fair Use[4] permits limited use of copyrighted material without acquiring permission from the rights holders. Similarly, in the UK, a recent Exception[5] (Feb 2016) to the copyright law allows researchers to make copies of copyright material for computational analysis. According to this Exception researchers are given the "Ability to mine all types of content/data". More specifically, "The exception permits any published and unpublished in-copyright works to be copied for the purpose of text mining for non-commercial research. This includes sound, film/video, artistic works, tables and databases, as well as data and text, as long as the researcher has lawful access.". This exception explicitly regulates a lawful behaviour and it is very convenient for researchers.

According to the International Comparative Legal Guides website, in Sweden big data and analytics are permitted[6].

Unfortunately, to date, in many countries, text and data mining copyright regulations remain implicit rather than explicit. There are practices though that help researchers. One

[3] Somehow related to this topic is the recently funded project in the UK: "Text-based measures of information quality in online health information".

[4] See https://en.wikipedia.org/wiki/Fair_use

[5] See https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception

[6] See https://iclg.com/practice-areas/data-protection/data-protection-2017/sweden#chaptercontent12

TABLE II
CORPUS STATISTICS

	# initial seeds	# retrieved seeds	# bootCatted URLs	URLs per seed: Mean	URL per seed: Median	URL per seed: SD	# words
Unigr.	13	13	112	8.61	9.3	3.57	91 118
Bigrams	215	142	689	4.85	4	3.16	618 491
Total	228	155	801	5.16	5	3.35	709 609

practice that as been adopted in some contexts is to scramble the content; another approach is to limit the use of the content to a certain number of characters; other practices are described in portals and forums[7].

As a matter of fact, texts in *eCare_Sv_01* have not been reproduced in their integrity. When BootCat retrieves and downloads a document, it automatically removes boilerplate and other parts of the original web documents. These cleaning procedures facilitate the use of the corpus for automatic text analysis. In practice, this means that some parts of the original web pages have been stripped out from the web documents stored in *eCare_Sv_01* when BootCat preprocessed the documents for the download.

Research-wise, working on a corpus and being unable to share it to allow experimental replication or contrastive analysis is not only frustrating but it also curtails future progress[8]. Since we wish to enlarge *eCare_Sv_01* over time via collective collaboration, *eCare_Sv_01* is made public and is freely available for research purposes. We are ready to remove any text(s) from the corpus upon an objection from its copyright holder, although to our knowledge, nobody has ever requested to remove any web text from collections crawled from the web, neither within the "web as a corpus" experience, nor within the "wacky" initiative, nor with Common Crawl corpus[9]. *eCare_Sv_01* is distributed under the following disclaimer: "Copyright is held by the author/owner(s) of the web documents included in the corpus. The documents in the corpus can be used for research purposes ONLY.".

VI. Human Annotation and Inter-rater Agreement

The annotation of documents in the corpus as being *lay* or *specialized* was carried out by a native speaker who participates in the project. The lay annotator works in Language Technology and has little knowledge of medical terminology.

To have an idea of the agreement between a lay annotator and an expert annotator, we asked a second annotator who works in Health Informatics to annotate a small sample out of the whole corpus. Then we measured the agreement between the two annotators.

Several inter-rater agreement measures exist [20]. All the inter-rater agreement measures have their strong points and their drawbacks and the use of one over the other depends on the data, the task and the situation. In our case, we wish to measure to what extent two members belonging to two different user groups (i.e. the lay and the expert) spontaneously agree when assessing the difficulty of medical language. Our expectation is that a lay person tends to label as "specialized" a larger number of medical documents than an expert person, who, conversely, tends to see as "lay" many documents that laypeople would consider to be "specialized". In order to test this assumption, we measured the inter-rater agreement by using the classic unweighted Cohen's *kappa* [21] and Krippendorff's *alpha* [22] to get a straightforward indication of the raters' tendencies. Cohen's κ assumes independence of the two coders and is based on the assumption that "if coders were operating by chance alone, we would get a separate distribution for each coder" [20]. This assumption intuitively fits our expectations. Krippendorff's α is similar to Cohen's κ, but it also takes into account the extent and the degree of disagreement between raters [20].

Table III shows the interrater agreement on the annotated texts. Interestingly, annotators tend to disagree more on documents harvested with unigrams (Row 1), while they agree more on documents harvested with bigrams (Row 2). All in all, both κ and α scores are approx. 0.5, and both these values indicate a "moderate" agreement according to the magnitude scale for κ [23], and the α range [24]. These values endorse our hypothesis that there exists a "user group bias". If we contextualize the results, this finding means that patients (who usually have a "lay" perspective) tend to perceive many documents as "specialized", while doctors would assess these documents simply "normal". This has a linguistic implication that affects LT applications in the eHealth field as a whole, and we encourage more in-depth investigation about this topic in the future.

TABLE III
INTER-RATER AGREEMENT VALUES

# documents	Percentage	Cohenś Kappa	Krippendorffś Alpha
112 (unigr. seeds)	75.9	0.52	0.51
236 (bigr. seeds)	82.2	0.60	0.60
348 (all)	80.2	0.57	0.57

[7] For instance, see http://linguistics.stackexchange.com/questions/9232/do-i-have-copyright-issues-when-making-a-corpus-from-the-web

[8] On this topic see also: Branco, A., Cohen, K.B., Vossen, P. et al."Replicability and reproducibility of research results for human language technology: introducing an LRE special section" Lang Resources & Evaluation 2017 51

[9] See http://commoncrawl.org/the-data/

VII. Supervised Learning: the Lay Perspective

In this section, we present two experiments based on lay-specialized text classification. We apply fully-supervised machine learning methods to explore how well supervised algorithms learn the labels applied by the *lay annotator*.

Experiment 1 focuses on scalability, and help us understand whether the size of the corpus has an impact of the classification results. In Experiment 2, we explore to what extent lay-specialized text classification is affected by noisy documents.

In these experiments, we relied on the Weka Machine Learning Workbench [25] (Explorer and Experimenter interfaces).

A. Quick-and-Dirty: Features and Noisy Texts

The first question to answer when performing lay-specialized text classification is: which features are most appropriate to represent lay and specialized medical sublanguages? Intuitively, one would argue that readability assessment features could well represent the difference between lay and specialized texts. A stable set of readability assessment features is available for Swedish and has been applied to several standard corpora [26]. Unfortunately, texts crawled from the web are noisy, also after being automatically cleaned by BootCat. For instance, texts may contain informal language (e.g. sv: "nå'n annan som hatar utredningen?" English: "somebody else who hates the investigation"), and unpredictable combinations of English words (e.g. "therapycounseling") are numerous. This means that the automatic extraction of readability assessment features from *eCare_Sv_01* would imply a regularization of the corpus that we have not planned yet. At this stage, we focus on how to leverage on noisy texts rather than on how to regularize them. For this reason, we decided to apply a filter that requires no text pre-processing, namely the *StringToWordVector* filter that converts strings (i.e. textual content) to vectors of words. Only two attributes were declared, namely *the textual content of the document* defined as "string", and the *sublanguage label* (either "lay" or "specialized") defined as "nominal".

B. Experiment 1: Lay-Specialized Text Classification and Scalability

We converted four subsets of the whole corpus into four datasets. The first dataset contains 156 documents; the second one 220 documents; the third one 337 documents; the fourth datasets includes the whole corpus and contains 801 documents. The four datasets contain some overlapping data since we wish to simulate the progressive expansion of the corpus over time by appending more documents to the original corpus. The rationale of this experimental setting is to observe whether and to what extent the performance of the classifiers deteriorates when increasing the corpus size.

Since we did not know in advance which type of machine learning modelling would be more suitable for this kind of data, we applied three standard algorithms that have very different inductive biases, namely *Decision Trees*, *Naive Bayes* and *SVM*. We used Weka's implementations of the these algorithms, i.e. *J48*, *Naive Bayes* and *SMO*. All the algorithms were run with standard parameters. We ran each of the algorithms via a metaclassifier (i.e. Classify - Meta - FilteredClassifiers) and we selected in turn each of the pre-decided classifiers together with the *StringToWordVector* filter (standard parameters). We applied 10-fold-crossvalidation. Results are shown in Tables IV, V, VI and VII (values have been truncated to two decimal places).

For the first dataset (156 documents), *J48* seems to be less suitable than *Naive Bayes* and *SMO*. *J48*'s k statistic is low, indicating that most of the corrected classifications happen by chance. The confusion matrix for J48 shows that lay texts are quite confusing for this classifier (only 48 TP vs 35 missclassified cases), while specialized texts are more clearly set apart (110 TP vs 27 misclassifications). *Naive Bayes* and *SMO* do a better job on this dataset: their averaged ROC area values are much higher than 0.5 (0.5 would mean that a classifier is random). On the second dataset (220 documents), *J48*'s performance values are equivalent to *Naive Bayes*'s and *SMO*'s. On the third dataset (337 documents), *SMO* shows better figures. The performance on the fourth dataset is similar to the third dataset.

In order to compare the performance of the three classifiers on the four datasets, we applied the Corrected Paired T-Test (two tailed) provided by Weka's Experimenter interface. Statistical significance was measured on the results of the three classifiers per dataset, and on the performance of each classifier on the four datasets. Statistical significance was measured at significance level of $P < 0.001$ on the weighted averaged F-measure. The test did not detect any statistically significant variation. We interpret these findings as a sign of stability since results show the robustness of the models to scalability issues. This experiment supports our claim that a corpus can be extended without causing any deterioration of the performance of LT applications.

C. Experiment 2: Lay-Specialized Text Classification With and Without Noise

In Experiment 2 we explored whether there exists a performance gap between text classification models trained on a collection containing noisy documents and text classification models trained on a collection containing only noise-less documents.

Results are shown in Table VII and Table VIII respectively. In order to compare the two sets of results, we measured the performance of the same algorithm on the two datasets. As in Experiment 1, statistical significance was measured at significance level of $P < 0.001$ on the weighted averaged F-measure. The test did not detect any statistically significant variation. We interpret these findings as a sign of resistance to noise in the lay-specialized text classification task. This experiment supports our claim that noise does not always negatively affect classification performance.

D. Discussion

Experimental results show that lay-specialized classification performance is good (averaged F-measure is above 0.70 in

TABLE IV
DATASET 1: 156 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
J48	0.14	62.8	0.62	0.62	0.62	0.63	0.62	0.42
NB	0.46	75.6	0.77	0.75	0.76	0.80	0.75	0.26
SMO	0.43	75.6	0.75	0.75	0.75	0.71	0.75	0.32

TABLE V
DATASET 2: 220 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
J48	0.38	71.8	0.71	0.71	0.71	0.69	0.71	0.33
NB	0.45	72.7	0.75	0.72	0.73	0.78	0.72	0.25
SMO	0.36	70.9	0.70	0.70	0.70	0.67	0.70	0.35

TABLE VI
DATASET 3: 337 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
J48	0.38	72.1	0.71	0.72	0.71	0.71	0.72	0.33
NB	0.46	73.5	0.76	0.73	0.74	0.80	0.73	0.23
SMO	0.50	77.1	0.77	0.77	0.77	0.75	0.77	0.27

TABLE VII
DATASET 4: ALL 801 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
J48	0.38	74.15	0.74	0.74	0.74	0.66	0.74	0.37
NB	0.45	73.9	0.78	0,73	0.74	0.83	0.73	0.23
SMO	0.49	78.6	0.78	0.78	0.78	0.74	0.78	0.29

most cases) and stable across classifiers and across datasets of different sizes.

In our view these results are promising for two main reasons. The first reason is *scalability*: Experiment 1 shows that results are essentially equivalent across samples of different sizes since we observe no statistically significant degeneration in the performance when scaling out. This is reassuring: we can imagine a scenario where we design a dynamic and extensible corpus whose size can be increased over time, and this will not affect the expectation of efficiency and reliability of LT applications when scaling out.

The second reason is *resilience to noise*: removing noisy documents from a corpus can be prohibitive in some contexts. Arguably, not all LT applications require high quality texts to ensure a good performance and reliable results, as we have shown in Experiment 2.

VIII. CONCLUSIONS AND FUTURE WORK

In this position paper we argued that 1) leveraging on a dynamic and extensible corpus does not necessarily imply scalability issues for LT applications; 2) leveraging on a noisy corpus does not necessarily imply decreases in performance. To support our claims we presented the results of two experiments in lay-specialized text classification using standard algorithms with standard parameters. Results are not only promising but also encouraging because we expect that more customized algorithms and optimized parameters can improve on the current performance of the classification models.

The paper presents several novelties. The first novelty is the creation of a very specialized web corpus with highly technical terms coming from SNOMED CT (Swedish version). This design is new since, to our knowledge, normally medical web corpora are built using documents related to common diseases (like varicella, measles, etc.) rather than to very specific illnesses.

We introduced the notion of "user group bias", which indicates that lay annotator and the expert annotator tend to disagree when asked to assess whether a document is lay or specialized. Our experience shows that the annotators' judgment is biassed towards their own expertise (or lack of expertise) in the medical field. This a new type of awareness that it is worth discussing in future.

Promising findings have been presented about corpus scalability and noise resilience. Corpus scalability implies that a corpus can be increased over time and this will not necessarily affect the performance of LT applications based on that corpus. Noise resilience indicates that it is not always necessary to remove noisy documents from a corpus to get reliable performance. Building LT applications that are resistant to noise is an important future direction in Language Technology. Another LT application that may remain unaffected by the noise-ness of corpus is automatic lexicon induction based

TABLE VIII
DATASET WITHOUT NOISY TEXTS: 462 DOCUMENTS

	k	Acc.	Avg. P	Avg. R	Avg. F	ROC A.	Avg. TP	Avg. FP
J48	0.36	72.29	0.72	0.72	0.72	0.69	0.72	0.35
NB	0.57	79.22	0.82	0.79	0,79	0.88	0,79	0.16
SMO	0.57	80.95	0.81	0.81	0.81	0.78	0.81	0.23

on distributional semantics, where the emphasis is on the contextual similarity rather than on the quality of writing style. We will test this assumption in future experiments.

Currently we are working on the definition of statistical measures that help us gauge the degree of domain-specificity (i.e. the "domainhood") of a corpus with respect to a general-purpose corpus.

Future work includes the expansion of the corpus with texts in other languages and related to diseases not necessarily classified as chronic in SNOMED CT, e.g. "tachycardia" or "dementia". Additionally, since the current version of *eCare_Sv01* is small, we plan to expand it also by relying on semi-supervised and weakly supervised learning.

ACKNOWLEDGMENTS

Thanks to all the reviewers of the manuscript who, with their feedback, helped improve the original draft. Thanks to Wiktor Strandqvist who contributed to the experiments with R scripts. Corpus, scripts and the output of the classification models are available from the project website: http://ecareathome.se. We encourage the replication of the results presented in this paper, and welcome improvements and further discussion. This research was funded by the distributed environment Ecare@Home funded by the Swedish Knowledge Foundation.

REFERENCES

[1] M. Alirezaie, H. Karl, and B. Eva, “A pattern language for smart home applications,” *Semantic Web*, vol. 00, no. 00, p. 00, 2017.

[2] M. Alirezaie, “Bridging the semantic gap between sensor data and ontological knowledge,” Ph.D. dissertation, Örebro university, 2015.

[3] L. Deléger, B. Cartoni, and P. Zweigenbaum, “Paraphrase detection in monolingual specialized/lay corpora,” *Building and Using Comparable Corpora*, 2013.

[4] M. Seedor, K. J. Peterson, L. A. Nelsen, C. Cocos, J. B. McCormick, C. G. Chute, and J. Pathak, “Incorporating expert terminology and disease risk factors into consumer health vocabularies,” in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 2013, p. 421.

[5] L. Deléger and P. Zweigenbaum, “Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora,” in *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*. Association for Computational Linguistics, 2009, pp. 2–10.

[6] K. F. Heppin, “Resolving power of search keys in medeval a swedish medical test collection with user groups: Doctors and patients,” Ph.D. dissertation, Ph. D. thesis, University of Gothenburg, 2010.

[7] E. Abrahamsson, T. Forni, M. Skeppstedt, and M. Kvist, “Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language,” in *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, 2014, pp. 57–65.

[8] V. Haslerud and A.-B. Stenström, “The bergen corpus of london teenager language (colt),” *Spoken English on computer*, pp. 235–42, 1995.

[9] R. Basili, M. T. Pazienza, and P. Velardi, “Acquisition of selectional patterns in sublanguages,” *Machine Translation*, vol. 8, no. 3, pp. 175–201, 1993.

[10] G. Grigonytė, M. Kvist, M. Wirén, S. Velupillai, and A. Henriksson, “Swedification patterns of latin and greek affixes in clinical text,” *Nordic Journal of Linguistics*, vol. 39, no. 01, pp. 5–37, 2016.

[11] M. Nyström, M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson, and H. Åhlfeldt, “Creating a medical english-swedish dictionary using interactive word alignment,” *BMC medical informatics and decision making*, vol. 6, no. 1, p. 35, 2006.

[12] M. Nyström, M. Merkel, H. Petersson, and H. Åhlfeldt, “Creating a medical dictionary using word alignment: the influence of sources and resources,” *BMC medical informatics and decision making*, vol. 7, no. 1, p. 37, 2007.

[13] N. Elhadad and K. Sutaria, “Mining a lexicon of technical terms and lay equivalents,” in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, 2007, pp. 49–56.

[14] V. V. Vydiswaran, Q. Mei, D. A. Hanauer, and K. Zheng, “Mining consumer health vocabulary from community-generated text,” in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 1150.

[15] D. Kokkinakis, “The journal of the swedish medical association-a corpus resource for biomedical text mining in swedish,” in *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey*, 2012.

[16] H. Dalianis, M. Hassel, and S. Velupillai, “The stockholm epr corpus–characteristics and some initial findings,” *Women*, vol. 219, no. 906, p. 54, 2009.

[17] H. Dalianis, A. Henriksson, M. Kvist, S. Velupillai, and R. Weegar, “Health bank-a workbench for data science applications in healthcare.” in *CAiSE Industry Track*, 2015, pp. 1–18.

[18] M. Baroni and S. Bernardini, “Bootcat: Bootstrapping corpora and terms from the web.” in *LREC*, 2004.

[19] V. Volansky, N. Ordan, and S. Wintner, “On the features of translationese,” *Digital Scholarship in the Humanities*, vol. 30, no. 1, pp. 98–118, 2015.

[20] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[21] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[22] K. Krippendorff, “Content analysis. beverly hills,” *California: Sage Publications*, vol. 7, pp. l–84, 1980.

[23] J. Sim and C. C. Wright, “The kappa statistic in reliability studies: use, interpretation, and sample size requirements,” *Physical therapy*, vol. 85, no. 3, p. 257, 2005.

[24] K. Krippendorff, “Computing krippendorff’s alpha-reliability,” 2011. [Online]. Available: http://repository.upenn.edu/asc_papers/43

[25] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[26] J. Falkenjack, K. H. Mühlenbock, and A. Jönsson, “Features indicating readability in swedish text,” in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, no. 085. Linköping University Electronic Press, 2013, pp. 27–40.