Canary: An NLP Platform for Clinicians and Researchers

Shervin Malmasi[1]; Nicolae L. Sandor[2]; Naoshi Hosomura[1]; Matt Goldberg[3]; Stephen Skentzos[4]; Alexander Turchin[1]

[1]Brigham and Women's Hospital, Boston, MA;
[2]Boston University, Boston, MA;
[3]Harvard University, Boston, MA;
[4]Kiip, Inc., San Francisco, CA

Keywords

Information extraction, clinical informatics, natural language processing

Summary

Information Extraction methods can help discover critical knowledge buried in the vast repositories of unstructured clinical data. However, these methods are underutilized in clinical research, potentially due to the absence of free software geared towards clinicians with little technical expertise. The skills required for developing/using such software constitute a major barrier for medical researchers wishing to employ these methods. To address this, we have developed Canary, a free and open-source solution designed for users without natural language processing (NLP) or software engineering experience. It was designed to be fast and work out of the box via a user-friendly graphical interface.

Correspondence to:
Alexander Turchin, MD, MS
Brigham and Women's Hospital, Boston, MA
Email: aturchin@bwh.harvard.edu

Appl Clin Inform 2017; 8: 447–453
https://doi.org/10.4338/ACI-2017-01-IE-0018
received: January 17, 2017
accepted: February 22, 2017
published: May 3, 2017
Citation: Malmasi S, Sandor NL, Hosomura N, Goldberg M, Skentzos S, Turchin A. Canary: an NLP platform for clinicians and researchers. Appl Clin Inform 2017; 8: 447–453
https://doi.org/10.4338/ACI-2017-01-IE-0018

1. Introduction

The adoption of Electronic Health Records (EHR) in the last two decades has greatly increased the amount of digital data available to researchers. Investigators require specific information stored in the EHR that relates to their research aims, such as drug dosage information or test results. While some of this information is readily available as structured data, a large part of it is stored in unstructured format, such as free-text provider notes. However, this natural language data is not directly usable for quantitative analysis or predictive analytics without NLP.

Given that these documents contain vital clinical information, researchers have been developing computational methods to process and mine them for information of interest. This is often done via *Information Extraction* (IE), the task of identifying and extracting relevant fragments of text from a larger, unstructured document. This often relies on the use of *Natural Language Processing* (NLP) methods to facilitate the intelligent extraction of meaningful information from unstructured text. Once identified, these fragments can then be converted to structured form. Individual investigators have been developing such methods to extract the information that they require from their own data sources. Once extracted, this information is used for biomedical research, clinical decision support, evidence-based medicine or further processing.

Information Extraction has been studied for many decades, going as far back as the 1970s, and its importance continues to increase as the amount of unstructured data, particularly on the web, grows. A number of different methodologies are used to tackle this task. Rule-based methods have been the classical approach, while the recent ascent of machine learning methods has also led to the adoption of statistical methods. These include supervised classification approaches, using both discriminative and generative models as well as sequence labelling models.

While the advantages of machine learning methods for information extraction cannot be denied, they are not a panacea and their limitations should be acknowledged. One challenge is that this category of predictive models can be a "black box" whose face validity is difficult to ascertain and errors harder to correct. In our experience we have found this to be particularly problematic within the clinical domain as researchers often require precise information about the reasoning underlying the classification assigned to a data point. In this regard, rule-based models may be easier to understand and implement, especially for clinicians.

Another critical challenge is that developing predictive models, particularly for NLP, requires large amounts of training data. Preparing this annotated data can be expensive and time consuming as multiple specialists must often manually review thousands of documents.

Lastly, we must also note that this training data may be difficult to obtain for rare phenomena. It may be difficult to gather enough data to train a predictive model, even with the appropriate resources. In our experience, manual review of as many as 10,000 documents can barely yield 50 positive instances of the target phenomena. This is not sufficient to reliably train a machine learning model.

2. Canary

To address the lack of a free NLP platform that would combine user-friendly interface with clinical face validity and support for extraction of rare concepts, we have developed Canary. Canary is designed for processing documents to support the extraction of information from natural language text using user-defined grammars and lexicons. The software, which can be downloaded for free (http://canary.bwh.harvard.edu/), guides the user through the successive steps of building a rule-based language model for identification of a particular concept or a set of concepts.

The main steps in building the language model include:

1. A preprocessing component for performing text normalization and mapping of acronyms and synonyms; these transformations simplify the text matching process.
2. A vocabulary manager that allows design of a case-specific ontology through creation of user-specified semantic categories (word classes). This functionality plays an important role in supporting creation of high-fidelity language models beyond what's available through standard ontologies. This approach allows inclusion of common misspellings or words that would only be in-

cluded in this semantic category in the specific context of the concept being sought. For example, "call in" [a prescription to the pharmacy] can be considered a semantic equivalent of "prescribe" in the context of medication management, but would not be found in this category in a standard ontology, such as SNOMED.

3. Creation of grammar rules that define how these word classes can be combined to form target phrases.
4. Definition of specific conditions that must be met for information to be extracted. For example, users can specify the presence of one or more phrases, such as a medication class and adverse reaction, as an output condition.

▶Figure 1 shows the main Canary user interface and highlights the interface sections that correspond to steps 1–3 above, along with specific examples of preprocessing, vocabulary and phrase structure rules that can be created.

Canary platform includes several advanced functionalities that facilitate information extraction tasks but are not universally found in the existing NLP packages:

a) Support for extraction of concept-value pairs (e.g. blood pressure or left ventricular ejection fraction)
b) Unicode support that allows creation of language models for languages other than English (a demonstration of a language model in Korean is included in the *Canary* package)
c) Inclusion of unspecified words in the phrase structure description (if two concepts can be separated by a given number of arbitrary words). For example, when writing a rule to extract blood pressure readings, we may wish to allow up to two (or more) unspecified words between the blood pressure concept reference and the value. This would correctly process sentences such as "*His **blood pressure** yesterday was **125/85***".
d) Support for recognition of anaphora and cataphora through information extraction criteria that span multiple sentences.
e) Portability: language models created using *Canary* platform can be shared between researchers, facilitating intra- and inter-institutional collaboration.
f) Parallel processing: *Canary* can take advantage of multi-CPU machines to process large amounts of data. While the speed of processing depends on the complexity of the language model and specific CPU setup, in a recent test running on 12 cores of Xeon® E5520 2.27 GHz CPUs using a language model comprised of 284 phrase structures, *Canary* processed 2.8 GB of text at an average speed of 10.4 MB / minute.

The overall Canary work flow, illustrated in ▶Figure 2, is as follows. The first step is data collection and preparation. The current version of Canary supports input in text format, and so any data from databases or EMR systems must be exported in a text format, which is commonly supported. Once the documents are prepared, a subset can be split into development and test sets which are annotated by human experts for evaluation. These annotated texts are then used by content experts to develop the Canary model and evaluate its performance. This evaluation can be done with traditional statistical methods: the Canary output can be compared against the ground truth to calculate precision (positive predictive value / PPV) and recall (sensitivity) as performance metrics. Once model performance is satisfactory, it can be finalized and used to process unannotated texts. The resulting output can be analyzed to measure prevalence and other statistics.

3. Use Cases

The *Canary* NLP platform can be applied for a wide range of tasks, several of which we will highlight here. Earlier, internal versions of our tools – the precursor to the current *Canary* software – have been successfully used by researchers in our group to conduct large-scale epidemiologic investigations [1–3]. In particular, information extraction has proven to be a valuable tool for studying issues revolving around adverse reactions to medications as this information is often not stored/updated in structured form [3]. To this end, our software was used to conduct a seminal study on statin side effects [1].

In other instances, specific information – such as decline of medications by patients – is only available in narrative form. It is anecdotally known that patients frequently decline medications that are recommended by their healthcare providers. However, little systematic data are available on this phenomenon. It is not known how commonly patients decline medications and how frequently they ultimately receive medications they initially declined.

Insulin is thought to be one of the medications that are especially frequently declined by patients. Many are reluctant to start injectable medications; others express fear that "once you start insulin, you can't get off it". Studies show that patients whose diabetes is poorly controlled on oral medications take a very long time to be started on insulin [4] insulin decline by patients could be one of the reasons for this. However, data on insulin decline remains extremely limited. Here, our software is also being used to conduct one of the first empirical evaluations of insulin decline. In our preliminary analysis of over 16 million notes, prevalence of this phenomena has been less than 0.5%, highlighting its rarity. In a similar vein, the software can be used to study the decline by patients of other medication classes or medical procedures and surgical interventions.

One reason for the paucity of research in this area is that information on patients declining medications is not easily available. As these patients declined the medication before any prescription was written, no trail is generated in the data sources that are typically used to study medication prescribing, such as pharmacy insurance claims or EMR medication records. Instead, medication decline is primarily recorded in narrative notes, requiring labor-intensive chart review. NLP software like *Canary* holds great promise for allowing clinical researchers to access the valuable pieces of relevant information locked away among millions of unstructured health records.

Canary can also be used to extract other values and measurements from narrative documents, such as blood pressure values, vital signs or tumor sizes from radiology reports. The software contains a number of sample projects demonstrating some of the abovementioned tasks.

In sum, we posit that information extraction is a key medical informatics tool that could be used broadly in biomedical research and clinical operations. Availability of a free NLP platform geared towards clinicians and researchers can therefore be of significant benefit to the medical community.

Multiple Choice Question

When performing information extraction, in which scenario could a rule-based approach be more suitable than a machine learning approach?

a) The target information is numeric
b) Training data is small or not readily available (CORRECT ANSWER)
c) No gold standard data is available
d) The data requires extensive preprocessing

Protection of Human Subjects

All research projects involving analysis of narrative electronic medical record data by Canary were reviewed by Partners HealthCare Human Research Committee, and the requirement for written informed consent was waived.

Conflicts of Interest

There are no conflicts of interest.

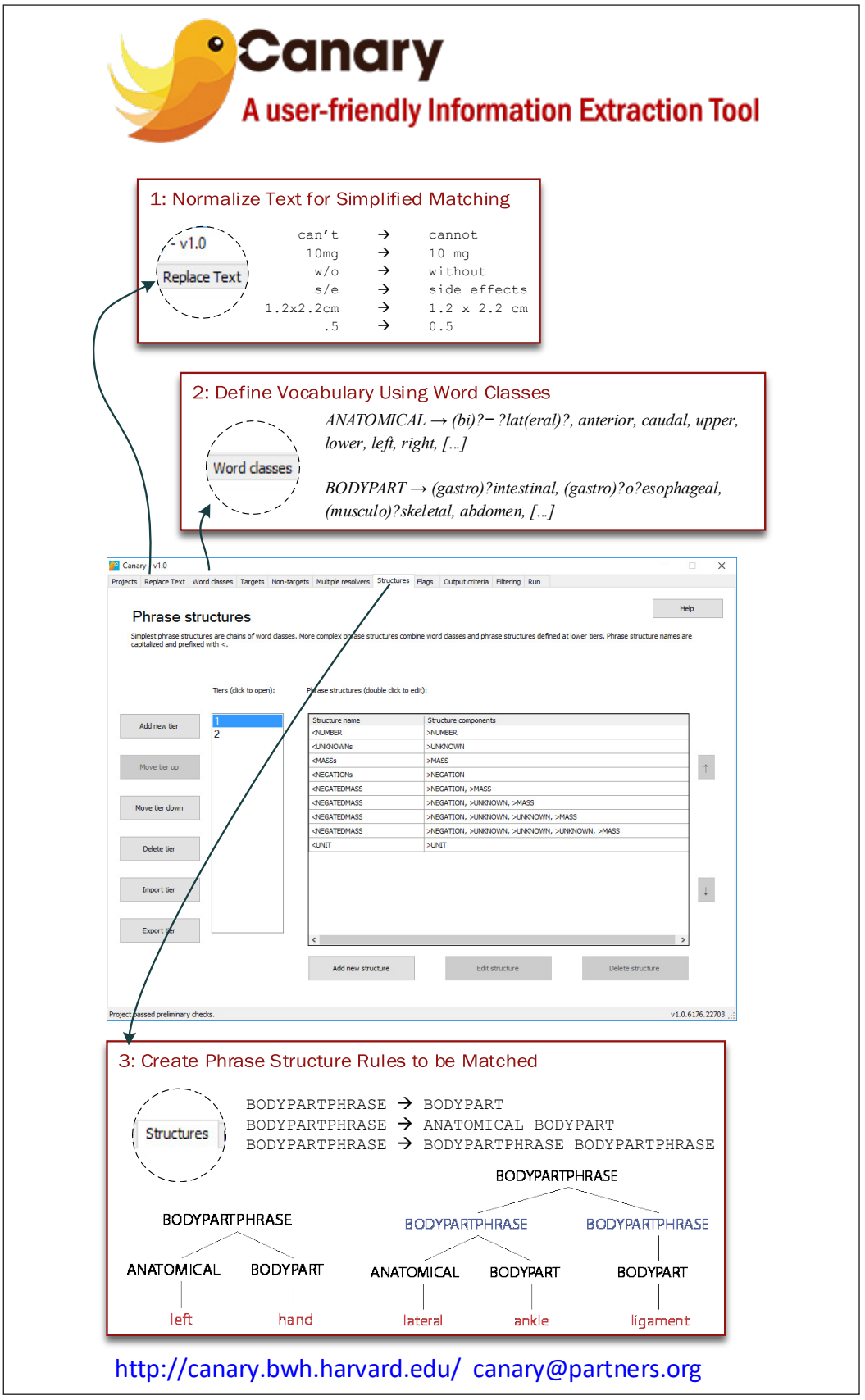

Fig. 1 An overview of the Canary software. The main Canary user interface is shown, along with examples of pre-processing, vocabulary and phrase structure rules that can be created.

)

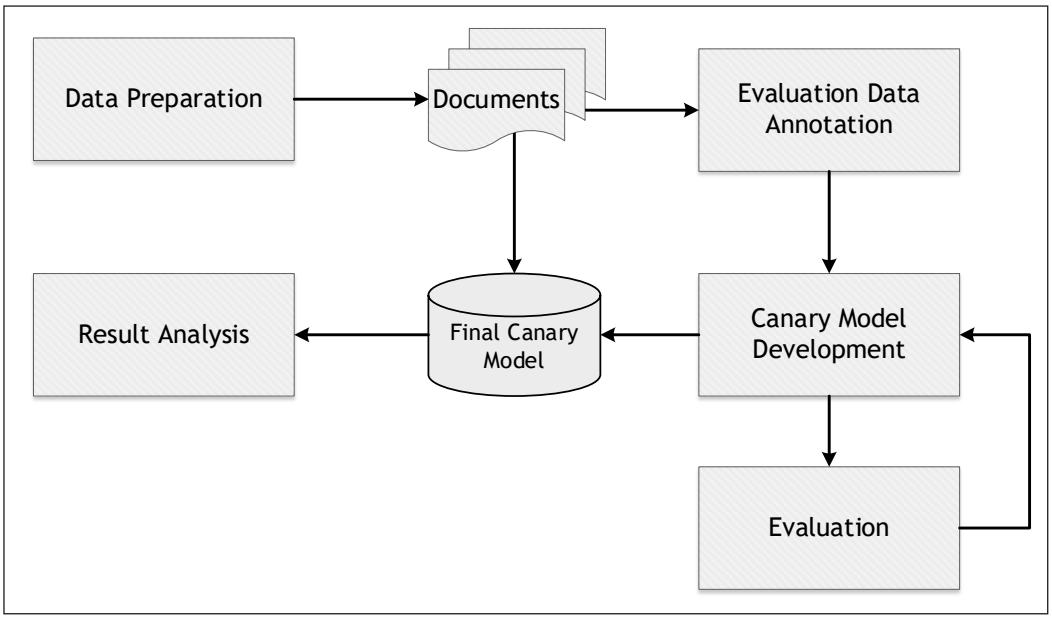

Fig. 2 A typical Canary project begins with data collection and annotation. The annotated data are used to develop and evaluate a model, which is then used to process unannotated documents.

References

1. Zhang H, Plutzky J, Skentzos S, Morrison F, Mar P, Shubina M, Turchin A. Discontinuation of statins in routine care settings: a cohort study. Annals of Internal Medicine 2013; 158(7): 526-534.
2. Zhang H, Plutzky J, Shubina M, Turchin A. Drivers of the Sex Disparity in Statin Therapy in Patients with Coronary Artery Disease: A Cohort Study. PLoS One 2016; 11(5): e0155228.
3. Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. Unstructured: Factors Affecting Adverse Drug Reaction Documentation in an EMR Repository. Proc AMIA Symp 2011: 1270-1279.
4. Rubino A, McQuay L, Gough S, Kvasz M, Tennis P. Delayed initiation of subcutaneous insulin therapy after failure of oral glucose lowering agents in patients with Type 2 diabetes: a population based analysis in the UK. Diabetic Medicine 2007; 24(12): 1412-1418.