MedicineAsk: An intelligent search facility for information about medicines

Ricardo Ferrão, Helena Galhardas, Luísa Coheur
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
ricardo.ferra@tecnico.ulisboa.pt
helena.galhardas@tecnico.ulisboa.pt luisa.coheur@inesc-id.pt

Abstract. Obtaining information quickly and easily is very important in the medical field. Natural Language Interfaces are one way to access this kind of information. MedicineAsk is a prototype that seeks to answer Portuguese Natural Language questions about medicines and active substances. It was designed to be easy to use so that questions may be posed by both medical staff and common users. Questions are answered through information previously extracted from the Infarmed’s Therapeutic Handbook and stored in a relational database. This document describes the extension of the Natural Language processing module of MedicineAsk. We focused on increasing the quantity of answerable user questions. First, we added machine learning techniques for question classification by using Support Vector Machines. Second, support for questions including anaphora and ellipsis has been implemented. We performed a validation over each of the new MedicineAsk features. Our improved MedicineAsk NLI answered 17% more questions than the previous version of MedicineAsk, with a further 5% increase when handling anaphora. We identified current limitations of MedicineAsk and suggested some solutions. This document also shows the state of the art on medical domain question answering systems.

Keywords: Natural Language, Medicine, Support Vector Machines, Anaphora Resolution.

1 Introduction

Every day, the data available on the Internet, in particular medical information, increases significantly. Medical staff and common users may have interest in accessing this information. Due to its nature, medical information often has to be accessed quickly. For example, a doctor may need to quickly access information in order to treat an emergency patient. A common user may have lost the information regarding one of his medicines and thus needs to urgently access the correct dosage for that medicine. For this reason, on-line medical information should be available through an interface that is fast and easy to use by most people.

There is a large amount of medical information of many different types and formats currently available on-line. This information is contained in either
databases of medicines and diseases or collections of papers containing important medical findings. Currently, to access this on-line information, users must either do it by hand (i.e., by manually reading a large volume of information and/or navigate through an index), learn a language to query the data (e.g., learn SQL to query a database with medical information), or use a keyword-based system.

One alternative to access medical information is through a Natural Language Interface (NLI). It has been shown that users often prefer Natural Language Interfaces over other methods such as keyword-based search (Kaufmann & Bernstein, 2007). While some NLIs have been developed for the medical field (Ben Abacha, 2012), they are still relatively new and none is available for the Portuguese language. This means that a Portuguese user who wants to access medical information available on-line must either use a system in a foreign language, which the user may not be fluent in, or use a traditional method, such as keyword-based search, like the one available in the Infarmed website\(^1\). Among various on-line services, Infarmed provides the *Prontuário Terapêutico*\(^2\) (Therapeutic Handbook), which publishes data about medicines and active substances approved to be sold in the Portuguese market. From hence forth, we refer to this information source as the “Infarmed website”. The Infarmed website contains information about medicines and active substances, such as their indications, adverse reactions, precautions, dosages and prices, among others. The user may access the information available on the Infarmed website by navigating an hierarchical index (which works similarly to the index of a book) or by using a keyword-based search. Navigating an index requires some knowledge of medical terms, and keyword-based search can provide incorrect or irrelevant results.

**MedicineAsk** is a question-answering prototype for information about medicines and active substances. This prototype was developed in the context of two master thesis (Bastos, 2009) (Mendes, 2011). MedicineAsk intends to solve the problems of the Infarmed website previously explained, by providing an NLI for the Infarmed website. The idea is that, by using an NLI, both common users and medical staff will be able to access the information on the Infarmed website in an easier and faster way. The MedicineAsk architecture is divided into two modules: Information Extraction and a Natural Language Interface.

The Information Extraction module is responsible for extracting information from the Infarmed website, processing it and inserting it into a relational database.

The Natural Language Interface enables users to access the information on the Infarmed website. Users can query about specific information regarding active substances and medicines, such as the price of a specific medicine or the indications of an active substance. The second version of MedicineAsk improved the first one since it was able to answer a larger number of questions.

The NLI of MedicineAsk still has limitations regarding what a user can ask the system. Questions that contain anaphora or ellipsis cannot be answered. In other words, if the interpretation of a question depends on a previous question

---

then MedicineAsk cannot answer that question. For example, in the questions “What are the indications of paracetamol? And what about the adverse reactions of that substance?” the second question does not directly mention the active substance. To answer the second question we must use the entity “paracetamol” located on the first question.

Furthermore, the previous version of MedicineAsk uses rule-based techniques to answer questions. While these techniques can achieve good results, they also require a user’s question to match a certain pattern, or to contain certain keywords. Machine learning techniques suffer less from these issues, and possibly achieve better results than the techniques used by the previous version of MedicineAsk.

In this paper we describe the main improvements brought to MedicineAsk, aimed at solving these issues.

This document is organized into five sections. Section 2 details the related work, in particular describes a question answering system in the field of medicine. Section 3 describes the improvements made to the NLI module of MedicineAsk in the scope of this document. Chapter 4 describes the validation performed on the new NLI module of MedicineAsk. Finally, Chapter 5 concludes and summarises future work.

2 Related Work

MEANS (Ben Abacha & Zweigenbaum, 2012) is a question-answering system in the medical domain. In a similar way to MedicineAsk, it creates a database, processes a question in natural language (in English), builds a query from that question and obtains the answer from the database. The database used is an RDF graph and the language used to query it is SPARQL. The database was created by extracting information from a medical corpus and annotating it in RDF. It analyses questions through rule-based and machine learning methods. The rule-based methods use MetaMap (Aronson, 2001), an online tool which finds and classifies concepts in text by mapping them to concepts from the United Medical Language System (UMLS). The machine learning version uses a combination of Conditional Random Fields (CRFs) to classify the concepts. MEANS supports questions about general medicine unlike MedicineAsk which focuses on questions regarding medicines.

Anaphora resolution consists on resolving an anaphora in a sentence by finding the corresponding antecedent, that is, the entity being referenced by the anaphora. Anaphora resolution is a topic that has been studied for many years. Hobbs’ algorithm (Hobbs, 1978) is one of the most known algorithms. Hobbs’ algorithm traverses the syntactic tree of the sentence the anaphora is located on, in a particular order, searching for possible antecedents using a breadth-first strategy. Antecedents that do not match the gender and number of the anaphor are not considered.

3 Improvements to the MedicineAsk NLI module

Our goal was to improve the MedicineAsk’s NLI module by increasing the number and variety of user questions that the system is able to answer. User questions can be too complex (e.g. contain useless information such as the name of the patient requiring a specific medicine) or too abstract (e.g. “What’s the price?”) to be answered in previous versions of MedicineAsk. While it is impossible to cover every possible question a user could pose, in this work we seek to cover enough questions so that the users of MedicineAsk can obtain the information they need by posing a Natural Language question in Portuguese.

3.1 Automatic question classification

The NLI module of the previous version of the MedicineAsk system uses rule-based and keyword spotting techniques (Galhardas, Coheur, & Mendes, 2012). Rule-based methods require a user’s question to exactly match a certain pattern. For this reason, these methods usually have low flexibility. Keyword spotting techniques require certain keywords to be present in the question, as well as dictionaries of those keywords to be built. In this work, we integrated a machine learning approach using Support Vector Machines (SVMs) (Zhang, 2003).

In order to understand how SVM solves questions we can compare all three available techniques of MedicineAsk. The rule-based technique answers questions by matching a question to a pattern. Each pattern is associated with a question type. Each word that is not part of the pattern is part of the named entities. The keyword spotting technique analyses the words in the question and, depending on the type of words present, determines the question type (e.g., if the word indications is present then it is a question about indications). It uses the same method to discover any named entities in the question. SVM has been trained so that each question type is mapped to a class. SVM then attempts to classify each user question into one of these classes, thus determining the question type. The named entities are discovered through a dictionary based annotator.

To answer a question using SVMs, MedicineAsk must use LUP (Reis Mota, 2012). LUP is a platform that can be used to run different Natural Language Understanding techniques (including SVMs) and compare the results of those techniques. Using the model corpus and dictionary of named medical entities, LUP uses SVM to determine the question type and the named entities present in the question. This information is then sent back to MedicineAsk, which can now build an SQL query based on the question type and named entities of the user question.

We adopted three different strategies in combining question classification techniques in MedicineAsk, in order to determine which method was more effective. These strategies use the currently available NLP techniques in MedicineAsk sequentially - rule-based, keyword spotting and SVM. We considered the following strategies to integrate SVM into Medicine in order to answer questions:

- **Strategy 1**: Rule-based, falling back to keyword spotting if no match is found
– **Strategy 2**: Rule-based, falling back to SVM classification if no match is found
– **Strategy 3**: Rule-based, falling back to keyword spotting if no match is found, and then falling back to SVM classification if keyword spotting fails as well.

Strategy 1 was featured in the previous version of MedicineAsk. It tries to answer a question using rules. If no match is found, it falls back on the keyword spotting method.

Strategy 2 is similar to Strategy 1: MedicineAsk first tries to match a question to a rule-based method, and if no match is found then it falls back on SVM.

Finally, Strategy 3 (see Figure 1) attempts to use every available method. If a question does not match any patterns using the rule-based method, then the keyword spotting technique is used. If no question type can be determined with this technique then MedicineAsk attempts to answer it through the SVM method.

![Figure 1: Strategy 3.](image)

The goal of using these strategies is to evaluate the difference of answering questions using only a single technique versus answering questions using combinations of those techniques. We only considered a few combinations among all the possible ones between these techniques. The rule-based method should be first as it is the most reliable if the user matches a question exactly with a pattern. SVM is the last technique to be used because SVM never fails to assign a class to a question.

### 3.2 Anaphora and Ellipsis

This work aims at supporting questions featuring anaphora and ellipsis. We do not intend to give full support to these types of questions. The goal is to answer some simple questions that contain these special cases and, in the future, these
features will be expanded in order to support an even larger amount of questions. We focus in medical entity anaphora, where the anaphora is a medical entity, such as a medicine or active substance. For example, “What are the indications of paracetamol? And what about the adverse reactions of that substance?” the anaphor is “that substance” and the antecedent is “paracetamol”.

The strategy to resolve medical entity anaphora is as follows. The first part consists on analysing regular questions with no anaphora. If the question is successfully answered, the question’s question type and named entities are stored in a data structure which we will call Antecedent Storage. For the example question “What are the indications of paracetamol?” the question type “indications” and the entity ”Paracetamol” are stored in the Antecedent Storage. Only the information of the last two questions is stored, however this number is configurable. This is because the older the question we analyse the less weight we give to them to help us resolve the anaphora, as the chance of them still being relevant decreases over time (Hobbs, 1978).

Afterwards it is possible to analyse questions with medical entity anaphora. If a question is analysed and no entities are found a case of anaphora is detected. In this case, we send the information of the question with anaphora to the Anaphora Resolver. The Anaphora Resolver then looks at the information of the question with anaphora and compares it to the information in the Antecedent Storage. If a possible antecedent for the question with anaphora is found then we return that antecedent as the entity, which will be user to answer the question. For the question “What are the precautions of that active substance?”, MedicineAsk identifies that this question has no medical entities and thus we have a case of medical entity anaphora. The Anaphora Resolver is then in charge of finding a possible antecedent. From the previous example, the Antecedent Storage includes information about the active substance “paracetamol”. This entity is then used in conjunction with the question type of “precautions” which leads MedicineAsk to answer the question “What are the precautions of paracetamol?”

The user is alerted to the fact that an anaphora was detected and another entity was used, showing him/her which entity it was. This is to prevent accidents where a user either forgot or misspelled the medical entity he wanted to query about, leading the system to think it was dealing with anaphora. For example if the user asked about the indications of paracetamol and then asked about the adverse reactions of mizolastina, but misspelled mizolastina, then the user would incorrectly receive the adverse reactions of paracetamol as an answer.

This implementation was created with the goal of being extensible. The rules used to determine if a given antecedent is compatible with the current question’s anaphora are stored in an XML file. By editing this XML file, it is possible to easily add new rules. It is also possible to change all of the rules to fit a different environment. This makes it possible to use this anaphora resolver for other environments, with the only changes necessary being in the XML file. The restriction is that the new environment must be somewhat similar to MedicineAsk. This means that the new system must deal with question answering, and the question’s information should be a question type plus a list of detected entities.
4 Validation

This chapter describes the evaluation of the version of MedicineAsk produced as a result of this thesis, which aimed to improve the MedicineAsk NLI module. We performed various experiments to test how each of the new features of MedicineAsk improve upon the previous version. Section 4.1 details the experimental setup used for both tests. Section 4.2 tests different question answering strategies without anaphora resolution. These strategies are detailed on Section 3.1. Section 4.3 tests different question answering strategies with anaphora resolution.

4.1 Experimental Setup

We collected a test set, the questionnaire test set, to compare the rule-based with the SVM approach. To this end, an on-line questionnaire composed of 9 different scenarios was distributed over the internet, using Facebook. Each scenario consists of a description of a problem that is related to medicines (e.g. “John needs to know the adverse reactions of Eeralgan, what kind of question should he ask?”). The participants were invited to propose one or more (natural language) questions for each scenario. We collected questions from 61 users, 19 medical professionals, and 42 common users.

For the tests detailed in this section we selected a subset of the questionnaire test set which includes 31 users. This subset of the test corpus included a total of 322 questions divided into 9 scenarios. The questions were not pre-processed in any way. Any errors and typos present in the questions were not removed.

4.2 Testing different question answering strategies

We measure the percentage of correctly classified questions from Test Corpus B. Figure shows the results of this experiment. An answer is correctly classified if the NLI returns the expected answer through the MedicineAsk website.

Some of the errors from this experiment include:

- User misspelling a word, namely medical entities (e.g. “What is Effermalgan for?”, the correct spelling is Efferalgan);
- User omitting a medical entity (e.g. “what is this medicine for?”);
- User asked the wrong question (e.g. “What are the indications of Efferalgan?” when the scenario was about precautions);
- Presence of words in the medical entity dictionary that are too common (e.g. “disease”).

We see that the addition of SVM to the answering process of MedicineAsk NLI brings improvements. Strategy 2 and Strategy 3’s percentage of correctly classified questions improved on Strategy 1 by 15% and 17% respectively. Both methods show good results, but Strategy 3 has better results. The reason why

Strategy 3 is superior to Strategy 2 is because the keyword spotting technique can answer some questions that SVM fails to answer. For example the question about indications *Que doenças se trata com Efferalgan?* (“What diseases are treated with Efferalgan?”) is classified as a question about interactions by SVM, so Strategy 2 fails to answer this question. The keyword spotting technique, however, can successfully classify it as a question about indications and so Strategy 3 can successfully answer this question. On the other hand, any question that cannot be answered by the keyword spotting technique will be sent to SVM which will possibly answer the question correctly. This is how Strategy 3 triumphs over Strategy 2. For this reason we have decided to use Strategy 3 as the final strategy in MedicineAsk.

4.3 Testing anaphora resolution

To demonstrate anaphora resolution we performed the same experiment as the one from Section 4.2 with anaphora resolution. We compared the results of Strategies 1 and 3 from that experiment, with the results of the same strategies but with anaphora resolution.

Figure 3 shows the percentage of questions correctly classified for the questions in Test Corpus B. An answer is correctly classified if the NLI returns the expected answer through the MedicineAsk website.

We see that, in total, Strategy 3 with anaphora resolution correctly answered 5% more questions than regular Strategy 3. Most of the questions missing an entity were correctly classified. Strategy 3 with anaphora resolution still seems to show the best results, and will be the strategy used by this version of MedicineAsk.

Anaphora resolution for medical entities brings improvements to MedicineAsk. It can also serve as a tool to prevent spelling errors by users, as long as they were querying the same entity between questions. To avoid errors a warning is shown telling the user that no entity was detected and a previous one was used.
5 Conclusions

This document presents an improvement to the NLI module of MedicineAsk, a system that answers Portuguese Natural Language questions about active substances and medicines. The main objective of this work was to increase the quantity of user questions that MedicineAsk can answer. We also aimed to test several different configurations and strategies of the question answering techniques, to determine which ones brought better results.

Strategy 3 shows an increase of 17% in question answering during the experiment described in Section 4.2. From the experiment described in Section 4.3 we see that Strategy 3 with anaphora resolution answers 5% more questions than regular Strategy 3 and 18% more questions than regular Strategy 1 with anaphora resolution. Thus we can see that both SVM and anaphora resolution can bring improvements to the NLI module of MedicineAsk.

We have identified several limitations with the new version of MedicineAsk, such as lack of support for questions that cover more than one topic. For example, to answer the question “What are the indications and adverse reactions of paracetamol?”, the user would have to pose two different questions to obtain all the information he seeks. Another issue is user mistakes, as seen in the 5% increase in correctly answered questions just from using anaphora resolution. The rule-based method has techniques to solve misspelled terms. However these techniques only work because the rule-based method knows exactly where the medical entity should be, and thus can obtain the misspelled word and correct it in order to try and obtain an answer. It would be interesting to expand this error correction function to questions analysed with the keyword spotting technique and SVM. As mentioned in Section 1, medical information must sometimes be accessed quickly. However, the current MedicineAsk system was made for traditional web browser interfaces. In emergency situations it may not be possible to access a computer in time. A mobile application of MedicineAsk could solve this issue. Making the information on the Infarmed website available through a
portable device such as a smart phone would make this information much more accessible.

References


