A Machine Learning based Natural Language Interface for a database of medicines

Ricardo Ferrão, Helena Galhardas, Luísa Coheur

INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
ricardo.ferrao@tecnico.ulisboa.pt
helena.galhardas@tecnico.ulisboa.pt
luisa.coheur@inesc-id.pt

Abstract. Medical information needs to be efficiently retrieved by medical staff and common users. Most retrieval engines rely on search methods based on keywords or interfaces that enable to navigate through the index of a book or a website. MedicineAsk is a prototype that enables users to access information about medicines through a Natural Language Interface (NLI) in Portuguese. In this paper, we compare the existing rule-based NLI module of MedicineAsk against a machine learning based one. We show that the machine learning based NLI is a better alternative to the rule-based methods for most scenarios. This suggests the possibility of a hybrid technique to take advantage of both methods.

1 Introduction

Many interfaces to medical information require the user to know how this information is organized and, sometimes, to be a medicine expert. For instance, the Portuguese Infarmed website offers access to data about medicines through the *Prontuário Terapêutico* (Therapeutic Handbook) [1] either by navigating through an index or by keyword search.

A Natural Language Interface (NLI) is an alternative to this type of interfaces. In previous work we proposed the MedicineAsk prototype [3], which provides a NLI to the Infarmed website. MedicineAsk extracts information from this site and stores it in a relational database. Then, it is able to answer user's questions in Portuguese. Preliminary experiments showed that the MedicineAsk NLI provided a better user satisfaction and ease of use when compared to the Infarmed website [5].

The NLI module of MedicineAsk is rule-based. In this paper, we follow a machine learning approach.

2 MedicineAsk: rule-based vs machine learning approach

Given a user question in Portuguese, the NLI module of MedicineAsk interprets and translates it into an SQL query, which is posed to the relational database that provides the answer. Figure 1 shows an example question and its internal representation

[1] From this point forward we shall refer to Infarmed's *Prontuário Terapêutico* as "the Infarmed website".

in MedicineAsk. The current interpretation step relies on handcrafted rules and keyword spotting [4]. Each user question is tested against each rule and, if a match occurs, the question type is identified (for instance, a question about indications represents a question type and a question about adverse reactions represents another question type). Moreover, a dictionary-based named entity recognizer is responsible for extracting medical entities. When the question type and the question entities are obtained, another set of rules generates the corresponding SQL query. If no rule can be applied to a given question, a keyword spotting technique is used.

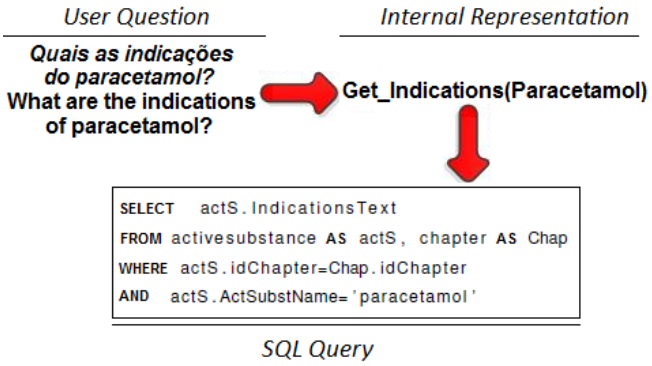

Fig. 1: Example of a user question and its internal representations in MedicineAsk

In this paper, we use Support Vector Machines (SVM) to find the question type in a one-versus-all strategy, as they led to state-of-art results in question classification [7]. To this end, we use LUP [6], a platform for Natural Language Understanding that includes the LIBSVM[2] implementation of SVMs, and supports several features, such as unigrams, bigrams and word shape.

3 Experiments

The training corpus was built from 450 questions previously collected [5]. These questions were divided into 15 question types. We collected a test set to compare the rule-based with the machine learning approach. To this end, an on-line questionnaire composed of 9 different scenarios was distributed over the internet. Each scenario consists of a description of a problem that is related to medicines (e.g. "John needs to know the adverse reactions of Efferalgan, what kind of question should he ask?"). The 58 participants were invited to propose one or more (natural language) questions for each scenario. In this preliminary experiment we used questions from 30 randomly chosen users, which included a total of 296 questions divided into 9 scenarios. We tested these questions against a rule- and a SVM-based NLI. Figure 2 shows the percentage

[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

of questions correctly classified, for each scenario (*1 to 9*) and for the the average of all scenarios (*Total*). For each scenario, we show the percentage of questions correctly classified for each feature set used by SVM and for the rule-based NLI.

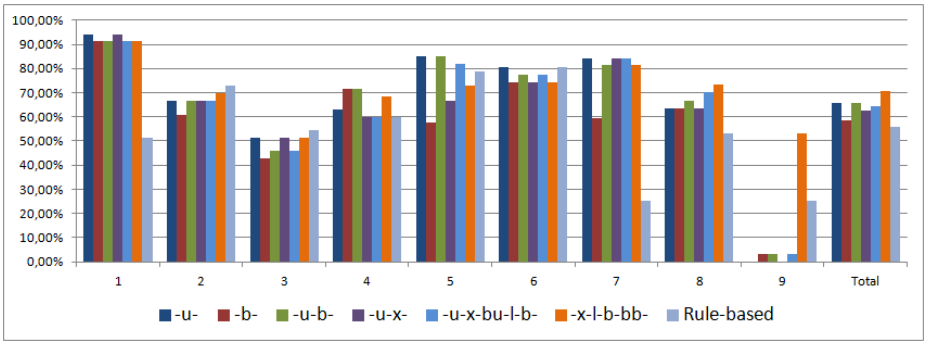

Fig. 2: Percentage of correctly classified questions by scenario for all users. The features are as follows: u - unigram, b - bigram, x - word shape, l - length, bu - binary unigrams, bb - binary bigrams.

Observing the total scores over all scenarios, we conclude that SVM has an advantage over the original rule-based method. This is due to machine learning methods being more flexible than rule-based ones. We can see that simply using unigrams can lead to very good results in most scenarios.

The majority of the cases in which the SVM failed were due to the fact that some words in the user's requests were not present in the corpus. For example, in the question *"Quais as doses pediátricas recomendadas da mizolastina?"* ("What are the recommended dosages for mizolastina?") we find that "doses", "pediátricas" and "recomendadas" are not present in the corpus. Also, some words were more frequently associated (in the training corpus) with a category different from the correct one. Scenario 9, which had the longest questions, got the worst results. Finally, none of the methods is robust to errors made by the user. For example, in some instances, the user misspelled certain words like medicine names.

4 Related Work

MEANS [2] is a question-answering system in the medical domain. Analogously 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. An ontology was defined in order to represent the concepts and relations used in MEANS. 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. MEANS analyses questions through rule-based and machine learning methods.

It uses manually built patterns to determine the question type. To recognize named entities, the rule-based methods use MetaMap [1], an online tool which finds and classifies concepts in text by mapping them to concepts from the Unified Medical Language System (UMLS). The machine learning method uses Conditional Random Fields (CRFs) to classify the medical concepts. MEANS supports questions about general medicine unlike MedicineAsk which focuses on questions regarding medicines.

5 Conclusions

The overall results of these preliminary tests show that SVM outperforms the rule-based methods. Currently, MedicineAsk tries to answer a question through a rule-based method, and if that fails it relies on keywords. We intend to replace the keyword method by an SVM and perform experiments to validate the gain obtained. We intend to enrich the corpus to further improve SVM's results. We also consider to use an ontology to represent named medical entities and other machine learning algorithms (e.g., CRFs) to find them. We also consider to use Portuguese morphosyntactical features or other linguistically motivated features to improve the recognition of questions.

ACKNOWLEDGEMENTS

This work was partially supported by national funds through FCT - *Fundação para a Ciência e a Tecnologia*, under the project PEst-OE/EEI/LA0021/2013 and the DataStorm Research Line of Excellency funding (EXCL/EEI-ESS/0257/2012).

References

1. Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. *AMIA Annu Symp Proc*, pages 17–21, 2001.
2. Asma Ben Abacha and Pierre Zweigenbaum. Medical question answering: Translating medical questions into SPARQL queries. *ACM SIGHIT International Health Informatics Symposium (IHI 2012)*, 2012.
3. Helena Galhardas, Vasco Duarte Mendes, and Luísa Coheur. Medicine.ask: a natural language search system for medicine information. *INFORUM 2012 - Simpósio de Informática*, 2012.
4. C. Jacquemin. *Spotting and discovering terms through natural language processing*. The MIT Press, 2001.
5. Vasco Duarte Mendes. Medicine.ask: an extraction and search system for medicine information. Master's thesis, Instituto Superior Técnico, 2011.
6. Pedro Mota, Lusa Coheur, Srgio dos Santos Lopes Curto, and Pedro Fialho. Natural language understanding: From laboratory predictions to real interactions. In *15th International Conference on Text, Speech and Dialogue (TSD)*, volume 7499 of *Lecture Notes in Artificial Intelligence*. Springer, September 2012.
7. J. Silva, L. Coheur, A. Mendes, and A. Wichert. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 2011.