Automatic Medical Knowledge Acquisition Using Question-Answering

Emilie PASCHE [a,1], Douglas TEODORO [a], Julien GOBEILL [a,b], Patrick RUCH [a,b]
Christian LOVIS [a]
[a] *Medical Informatics Service, University Hospitals of Geneva and University of Geneva, Switzerland*
[b] *College of Library Sciences, University of Applied Sciences, Geneva, Switzerland*

Abstract. We aim at proposing a rule generation approach to automatically acquire structured rules that can be used in decision support systems for drug prescription. We apply a question-answering engine to answer specific information requests. The rule generation is seen as an equation problem, where the factors are known items of the rule (e.g., an infectious disease, caused by a given bacteria) and solutions are answered by the engine (e.g., some antibiotics). A top precision of 0.64 is reported, which means, for about two third of the knowledge rules of the benchmark, one of the recommended antibiotic was automatically acquired by the rule generation method. These results suggest that a significant fraction of the medical knowledge can be obtained by such an automatic text mining approach.

Keywords. knowledge discovery, medical guidelines, infectious disease

1. Introduction

Drug prescription is an important source of error in medicine and antibiotic prescriptions have often been reported as non-compliant with good medical practices [1]. The main issues range from the large spectrum of available antibiotics, the time necessary to get the laboratory results and the absence of clear recommendations easily accessible at the point-of-care. Together with other societal causes (e.g., self medication, animal feeding), inappropriate antibiotic usage has been identified as a leading cause of bacterial resistance to antibiotics [2, 3]. As corollary, it is also responsible of an important increase of healthcare costs, hospitalization duration and adverse effects. Indeed, one of the main challenges when applying clinical decision support to operational settings either for population monitoring or patient care, is the acquisition of a reliable entity relation models [4]. Unlike historical expert systems [5], inspired by early work in artificial intelligence, we do not attempt to build a model using formal logic formalisms, but we design an original experiment to automatically extract rules from a corpus of legacy contents. Indeed, seminal works in clinical decision supports were massively dependent on domain-specific human-expertise to be maintained, which resulted in very limited (e.g., *bacterial meningitis*) and non-scalable applications of poor interest. As part of the DebugIT project, we investigated the ability to automatically generate the expert knowledge in a given domain.

[1] Corresponding Author: Emilie Pasche, University Hospitals of Geneva, Rue Gabrielle-Perret-Gentil 4, 1211 Geneva 14, Switzerland; E-mail: emilie.pasche@sim.hcuge.ch.

The matter of this report is the generation of machine-readable legacy knowledge rules, which can be used by clinical decision support systems to improve antibiotic usage [6]. The idea is to automatically translate clinical guidelines into logical rules to help care providers to perform their duties [7]. To assess our rule generation approach, a set of rules was manually crafted. Nevertheless, manual generation of medical theories is both labor intensive and time consuming. It requires a high level of expertise in various fields (medicine, medicinal chemistry, biology, etc) and a familiarity with ontology management. Although automatic generation of the domain knowledge at quality levels achieved manually is beyond the power of current technologies, it is argued that building large-scale biomedical knowledge bases cannot only be achieved using text mining methods, as suggested in [8] for molecular biology.

2. Data and Methods

Typical recommendations are useful to prescribe the most appropriate antibiotic, according to different parameters, patient situation, clinical assessment, but also costs, benefits, adverse effects, as well as the risk of resistance's development. As case study, we started investigating guidelines for the geriatrics (Table 1) and the surgery services from the University Hospitals of Geneva (HUG).

Table 1. Slightly simplified sample of the guidelines

Pathologies	Pathogenic agents	Antibiotics	Alternatives	Treatment duration
Cholecystitis	Enterobacteriaceae Enterococcus Clostridium sp	ceftriaxone 1 g/24h iv + metronidazole 500 mg/8h iv	amoxi./clav. 1,2 g/8h iv	10–14 days
Gastroenteritis	Campylobacter Salmonella E. coli Shigella (rare)	ciprofloxacin 500 mg/12h po	co-trimoxazole forte (160mg TMP/ 800mg SMX)/12h po & clarithromycin 500 mg/12h po	5–10 days

To transform such verbose documents into machine-readable data, the guidelines are transformed in database tuples. The translation from French to English was performed manually, assisted by French-to-English translation tools (e.g., http://eagl.unige.ch/EAGLm/), and a SNOMED categorizer [9]. For some of the queries, several answers were possible – three on average – as shown in Table 1, where *cholecystitis* caused by *enterococcus* can be treated by three different antibiotics: *ceftriaxone*, *metronidazole* and *amoxicillin-clavulanate*; each of them is unambiguously associated to a unique terminological identifier.

Automatically-generated rules are represented as triplets: 1) disease, 2) pathogen and 3) antibiotic. The objective of the automatic rule generation is to produce one of the items (so-called the *target*) of the triplet using the two other items (so-called the *sources*). The discovery of the third item relies on an advanced question-answering engine (EAGLi: Engine for Question Answering in Genomic Literature, http://eagl.unige.ch/EAGLi, [10]). Thus, the rule induction problem is reformulated as a question-answering problem: each information request is associated with a set of possible answers. Three types of questions are designed:

1. Drug: Which *antibiotic* should be used against *this pathogen* causing *this disease*?
2. Pathogen: Which *pathogen* is responsible for *this disease* treated by *this antibiotic*?
3. Disease: Which *disease* is caused by *this pathogen* and treated by *this antibiotic*?

Further, two search engines, corresponding to different search models were tested: easyIR (a relevance-driven search engine well known for outperforming other search methods on MEDLINE search tasks [11]) and PubMed (the NCBI's Boolean search instrument). All targets were normalized using standard terminologies: antibiotics and diseases in SNOMED CT or MeSH and bacteria in NEWT. To find the antibiotics, knowing the disease and the pathogen, a list of 72 antibiotic targets was defined, corresponding mostly to the UMLS Semantic Type T195. To find the bacterial pathogens, knowing the disease and the antibiotic, we suggested a subset of the NEWT terminology corresponding to the bacteria taxonomy. To find the pathological processes, knowing the pathogen and the antibiotic, we proposed a list of MeSH terms corresponding to disease, corresponding to the following UMLS Semantic Types T020, T190, T049, T019, T047, T050, T033, T037, T047, T191, T046 and T184. For the antibiotics category, we tested the use of both SNOMED CT and the MeSH. Synonyms from these terminologies were also evaluated. Thus, *amoxicillin with clavulanate potassium* can also be mentioned as *amoxicillin-clavulanic acid* or *augmentin*.

Furthermore, to fine-tune the question-answering module, several descriptors, in particular generic ones, needed to be removed. Thus, *infectious diseases* or *cross-infection* were removed from the descriptor list for the disease type of target. Finally, specific keywords were used to refine the search equation. Thus, we added context specific descriptors such as *geriatrics*, *elderly*, etc. The impact of general keywords was also tested such as *recommended antibiotic*, *antibiotherapy*, etc.

3. Results and Discussion

The evaluation of our results is done using TrecEval, a program developed to evaluate TREC (Text Retrieval Conferences) results using NIST (US National Institute of Standards and Technologies) evaluation procedures. As usual in information retrieval [12] and factoid question answering tasks, we focus on precision-oriented metrics. In particular, the so-called precision at recall 0, i.e., the precision of the top-returned answer, is used to evaluate the effectiveness of our approach. To complement this metrics, which provides the precision of the system used without user interaction, we also measure the recall of the system achieved by the top five answers. Thus, we try to estimate how useful such a system would be when used by an expert able to validate the ranked output of the guideline generator.

The fifty triplets generated manually are used as gold standard. Each rule/query concern a specific disease caused by a specific pathogen and is represented by a tuple of four columns. Diseases, pathogens and antibiotics were entered for each entry. Optionally, conditions were also added depending on the entries, such as weight or age. The fine-tuning of the engine has been done based on the TREC Genomics competitions: see e.g., [11]. Evaluation is done with a focus on the precision of the first retrieved antibiotic, which corresponds to the top-precision, noted P0 in the following.

As for the impact of the specific keywords, we observe that five keywords provided an important improvement compared to the baseline system: 1) *guidelines*; 2) *antibiotherapy*; 3) *recommended antibiotic*; 4) *aged*; and 5) *recommended antibiotic aged*. The best results were obtained with the expression *recommended antibiotic*. Top precision was improved by +20%. It is observed that using a relevance-driven ranking strategy (easyIR) returns answers, including relevant ones, for more questions

(coverage improved by 56%), than using a Boolean retrieval model (PubMed), which in contrast tends to provides a slightly better precision.

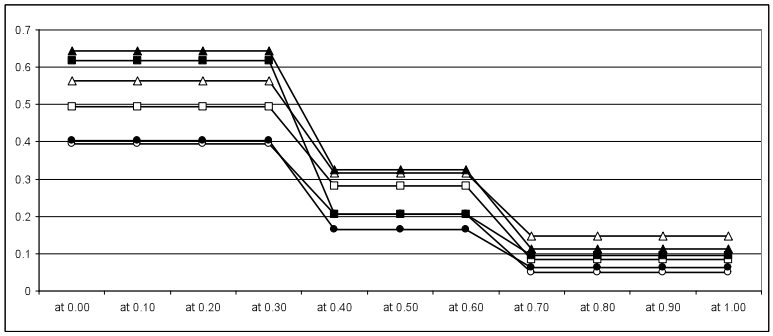

Figure 1. Comparative precision curves at various recall levels of easyIR, without antibiotic synonym and without keyword (○); PubMed, without antibiotic synonym and without keyword (●); easyIR, without antibiotic synonym and with the keyword *recommended antibiotic* (□); PubMed, without antibiotic synonym and with the keyword *recommended antibiotic* (■); easyIR, with antibiotic synonyms and with the keyword *recommended antibiotic* (Δ); PubMed, with antibiotic synonyms and with the keyword *recommended antibiotic* (▲).

The mean average precision (map = 0.36), which averages the precision over the different recall points, is also maximal with the Boolean engine, however half of the questions do not receive any answer because, as often with specific conjunctive Boolean queries, no documents are returned. When looking for a treatment considering a question containing a pair of {pathogen; pathology} (type #1), the system achieved a P0 of 64%, which means that about two times out of three, one of the recommended antibiotic is retrieved by the tool in the first position (Figure 1). The recall at five documents is of 48%, which means that about half of the recommended antibiotics are proposed in the top five retrieved antibiotics.

From the Figure 1, we also observe that the two engines, which tend to perform very similarly on average, seem not to behave similarly regarding their respective ranking power. While the Boolean engine outperforms the relevance-based model regarding precision at five documents, precision at lower ranks of the relevance-based engine is superior; thus suggesting that combining together the two engines could be beneficial to perform the task, as proposed for instance in [13].

Furthermore, a more detailed error and statistical analysis will be needed to separately establish significance of the different settings. Because guidelines edition is today a human-performed labor, conducted by domain specialists, comparison with existing practices is difficult. Nevertheless, it is well-known that inter-expert agreement is difficult to achieve – Funk and Reed [14] reports on a kappa ratio of about 60% for keyword assignment – therefore our study's underlying assumption that a theoretical precision of 100% is reachable should be also questioned.

4. Conclusion

In this report, we showed how text mining instruments such as question-answering engines can be used to automatically or semi-automatically generate medical expert knowledge. As proof of concept, we investigated the generation of {infectious disease; bacteria; antibiotics} association rules. Our approach was able to generate well-formed

rules for up to 64% of our infectious disease benchmark. It is thus expected that about two third of the domain knowledge predicted by our method is fully valid. Although partial, this result demonstrates that text-mining methods can automatically generate a significant subset of the medical expert knowledge, as available in a large text repository such as the MEDLINE digital library. As future work, we plan to use alternative corpora such as ClearingHouse (http://www.guidelines.gov) which can be used as valuable complement to MEDLINE for clinical guidelines and evidence-based medicine. Our approach should then be integrated into an interactive tool for creating and validating rules for drug prescriptions, which should facilitate antibiotic prescription rules management and guidelines edition in large healthcare institutions and public health administrations.

Acknowledgements. This experiment has been supported by the EU-IST-FP7 DebugIT project # 712139. The EAGLi question-answering framework has been developed thanks to the SNF Grant # 325230-120758.

References

[1] Mora, Y., Avila-Agüero, M.L., Umaña, M.A., Jiménez, A.L., París, M.M., Faingezicht, I. (2002) Epidemiological observations of the judicious use of antibiotics in a pediatric teaching hospital. *International Journal of Infectious Diseases* 6:74–77.

[2] Iosifidis, E., Antachopoulos, C., Tsivitanidou, M., Katragkou, A., Farmaki, E., Tsiakou, M., Kyriazi, T., Sofianou, D., Roilides, E. (2008) Differential correlation between rates of antimicrobial drug consumption and prevalence of antimicrobial resistance in a tertiary care hospital in Greece. *Infection Control and Hospital Epidemiology* 29:615–622.

[3] Gould, I.M. (2008) The epidemiology of antibiotic resistance. *International Journal of Antimicrobial Agents* 32(Suppl 1):S2–S9.

[4] Lovis, C., Colaert, D., Stroetmann, V.N. (2008) DebugIT for patient safety – improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. *Studies in Health Technology and Informatics* 136:641–646.

[5] Perry, C.A. (1990) Knowledge bases in medicine: A review. *Bulletin of the Medical Library Association* 78:271–282.

[6] Pestotnik, S.L., Classen, D.C., Evans, R.S., Burke, J.P. (1996) Implementing antibiotic practice guidelines through computer-assisted decision support: Clinical and financial outcomes. *Annals of Internal Medicine* 124:884–890.

[7] Harvey, K., Stewart, R., Hemming, M., Moulds, R. (1983) Use of antibiotic agents in a large teaching hospital. The impact of antibiotic guidelines. *Medical Journal of Australia* 2:217–221.

[8] Baumgartner Jr., W.A., Cohen, K.B., Hunter, L. (2008) An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *Journal of Biomedical Discovery and Collaboration* 3:1.

[9] Ruch, P., Gobeill, J., Lovis, C., Geissbühler, A. (2008) Automatic medical encoding with SNOMED categories. *BMC Medical Informatics and Decision Making* 8(Suppl 1):S6.

[10] Gobeill, J., Ehrler, F., Tbahriti, I., Ruch, P. (2007) Vocabulary-driven passage retrieval for question-answering in genomics. *TREC,* National Institute of Standards and Technology, http://trec.nist.gov/pubs/trec16/papers/uhosp-geneva.geo.final.pdf.

[11] Aronson, A.R., Demner-Fushman, D., Humphrey, S.M., Lin, J., Liu, H., Ruch, P., Ruiz, M.E., Smith, L.H., Tanabe, L.K., Wilbur, J. (2005) Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. *TREC,* National Institute of Standards and Technology, http://trec.nist.gov/pubs/trec14/papers/nlm-umd.geo.pdf.

[12] Singhal, A. (2001) Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24:35–43.

[13] Fox, E.A., Shaw, J.A. (1994) Combination of multiple searches. *TREC*, National Institute of Standards and Technology, http://trec.nist.gov/pubs/trec3/papers/vt.ps.gz.

[14] Funk, M.E., Reid, C.A. (1983) Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association* 71:176–183.