QA-driven Guidelines Generation for Bacteriotherapy

Emilie Pasche, PhD[1], Douglas Teodoro, PhD[1], Julien Gobeill, PhD[1,2], Patrick Ruch, PhD[1,2] and Christian Lovis, MD[1]
[1]Medical Informatics Service, University Hospitals of Geneva and University of Geneva,
[2]College of Library Sciences, University of Applied Sciences, Geneva, Switzerland

Abstract

PURPOSE: We propose a question-answering (QA) driven generation approach for automatic acquisition of structured rules that can be used in a knowledge authoring tool for antibiotic prescription guidelines management. METHODS: The rule generation is seen as a question-answering problem, where the parameters of the questions are known items of the rule (e.g. an infectious disease, caused by a given bacterium) and answers (e.g. some antibiotics) are obtained by a question-answering engine. RESULTS: When looking for a drug given a pathogen and a disease, top-precision of 0.55 is obtained by the combination of the Boolean engine (PubMed) and the relevance-driven engine (easyIR), which means that for more than half of our evaluation benchmark at least one of the recommended antibiotics was automatically acquired by the rule generation method. CONCLUSION: These results suggest that such an automatic text mining approach could provide a useful tool for guidelines management, by improving knowledge update and discovery.

Introduction

Antibiotic prescriptions have often been reported as non-compliant with good medical practice[1]. Indeed, the clinician empirically prescribes an antibiotic without or before any microbiological result, leading to the choice of broad-spectrum antibiotics but also to incorrect prescriptions. Overuse and misuse are among the main causes of antibiotic resistance of bacteria[2, 3]. As a corollary it is also responsible for a major increase in health care costs, hospitalization stays and adverse effects. Today, improving antibiotics use is clearly a public health priority.

Medical recommendations[4] are useful to prescribe the most appropriate antibiotic, according to different factors, such as patient situation, clinical assessment, but also costs, benefits, adverse effects and the risk of resistance development. Nevertheless, recommendations are often difficult to access at the point of care. Moreover, medicine is a moving target and is massively dependent on domain-specific human-expertise to be maintained. As part of the DebugIT project[5], we attempt to develop an interactive user-friendly tool for creating, editing and validating rules for antibiotic prescriptions, which should facilitate bacteriotherapy rules management and guidelines edition in large healthcare institutions and public health administrations.

Previously[6], we designed an original experiment for automatic extraction of rules from a corpus of legacy contents using simple search strategies, that can be used in clinical decision support systems to improve antibiotic usage[7]. The matter of this report is the description of further investigations of the generation of machine-readable legacy knowledge rules by developing more advances acquisition methods likely to combine together several heterogeneous search algorithms.

Data and Methods

Automatically-generated rules are defined as equations: an antibiotic A is prescribed under conditions B, C, etc. The objective of the automatic rule generation is to produce one of the items (the so-called *target*) of the equation using the other items (the so-called *sources*). The discovery of the target item relies on an advanced question-answering engine (EAGLi: Engine for Question Answering in Genomic Literature, http://eagl.unige.ch/EAGLi[8]). Thus, the rule induction problem is reformulated as a question-answering problem, with a ranked list of candidate answers as output. Different equations are proposed. They can be based on a combination of triplets such as {disease, pathogen, antibiotic}. As shown in the following questions:

1. What antibiotic A should be used against the pathogen P which causes the disease D?
2. What pathogen P is responsible for the disease D treated by the antibiotic A?
3. What disease D is caused by the pathogen P and treated by the antibiotic A?

Equations can also be based on pairs, such as {surgical procedure, antibiotic}. Two types of questions are designed:

1. What antibiotic A should be used as an antibioprophylaxis for the surgical procedure S?
2. What surgical procedure S requires the antibiotic A as an antibioprophylaxis?

In our experiments, we use two different search engines, corresponding to two different search

models: easyIR, a relevance-driven search engine well known for outperforming other search methods on MEDLINE search tasks[9], and PubMed, the NCBI's antichronological and Boolean search instrument. Finally, a combination of the outputs of the two engines (PubMed and easyIR) is tested to hopefully combine the power of both engines.

The set of possible answers (the so-called *target*) was defined based on terminologies, such as the MeSH or WHO-ATC. Previously we defined two antibiotic target sets. The *UMLS T195 target* consisted in a subset of the MeSH terminology, corresponding to the UMLS Semantic Type T195. The *WHO-ATC target* consisted in a subset of the WHO-ATC terminology, corresponding to a set of seventy antibiotics available at the University Hospitals of Geneva. This target includes only one term for describing each entity. A new antibiotic target set, the *hybrid target*, was defined, which merges together entities from the two terminologies. This set was limited to seventy possible answers. The MeSH terminology was used to define terms for each of the antibiotics, including synonymous terms. The WHO-ATC terminology was used to define identifiers, providing a classification of drugs according to their anatomical, therapeutic and chemical properties. The disease target set consisted of a list of MeSH terms corresponding to the following UMLS Semantic Types T019, T020, T033, T037, T046, T047, T049, T050, T184, T190 and T191.

Furthermore, fine tuning the question-answering modules was needed: it includes terminology pruning. Indeed, 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. Moreover, specific keywords were used to refine the search equation in order to retrieve more accurate results. Thus, we added context-specific descriptors such as *geriatrics*, *elderly*, for geriatric guidelines, etc. The impact of general keywords, such as *recommended antibiotic*, *antibiotherapy*, etc, is also tested.

The evaluation of our system was based on two sets of manually-generated rules extracted from the guidelines provided by the University Hospitals of Geneva (HUG). The first benchmark consisted in 64 triplets from guidelines designed for geriatric departments. Each rule/query concerned a specific disease caused by a specific pathogen and was represented by a tuple of four columns. Diseases and their corresponding MeSH identifiers, pathogens and their corresponding NEWT identifiers and antibiotics and their corresponding WHO-ATC identifiers were entered for each entry. Optionally, conditions were also added when available, such as the severity of the disease or the way it was acquired. An example of such a rule is provided in Table 1. For most of the queries several answers are possible – three on average – as shown in Table 1, where *severe diverticulitis* caused by *enterobacteriace* can be treated by three different antibiotics: *ceftriaxone*, *metronidazole* and *piperacillin-tazobactam combination*.

Disease	Pathogen	Antibiotics	Other condition
diverticulitis (D004238)	entero-bacteriaceae (543)	amoxicillin-potassium clavulanate combination (J01CR02); ciprofloxacin (J01MA02); metronidazole (J01XD01)	
diverticulitis (D004238)	entero-bacteriaceae (543)	ceftriaxone (J01DD04); metronidazole (J01XD01); piperacillin-tazobactam combination product (J01CR05)	severe

Table 1. Example of manually-generated rules for the first equation: terminological identifiers are provided in parenthesis. In this example, the infection can be treated by five different antibiotics. The use of *ceftriaxone* requires the condition *severe.*

Surgical operations	Antibiotics	Other conditions
laparotomy (D007813)	cefoxitin (J01DC01)	
cerebrospinal fluid shunt (D002557)	cefazolin (J01DB04); vancomycin (J01XA01)	external

Table 2. Example of manually-generated rules for the second equation: terminological identifiers are provided in parenthesis. In this example, infection during external cerebrospinal fluid shunt is prevented using *cefazolin* and *vancomycin*.

The second benchmark consisted of 25 pairs extracted from guidelines specialized for surgery. Each rule/query concerns a specific surgical procedure and is represented by a tuple of three columns. Surgical procedures and their corresponding MeSH identifiers as well as antibiotics and their corresponding WHO-ATC identifiers were entered for each entry. Optionally, a set of conditions was also added depending on the entries, such as

properties of the wound (clean, contaminated). An example of such a rule is provided in Table 2.

These guidelines were transformed manually from free text to database tuples. The translation and transformation process was assisted by tools, such as the French-to-English translation tool EAGLm (http://eagl.unige.ch/EAGLm/), or the SNOMED SNOCat categoriser[10], which helps assigning SNOMED CT descriptors to any textual content.

Results and Discussion

The evaluation of our results is done with TrecEval, a program developed to evaluate TREC (Text Retrieval Conferences) results using NIST (US National Institute of Standards and Technologies) evaluation procedures. Fine-tuning of the engine was based on the TREC Genomics competitions[9]. Two measures are selected to evaluate our results: the precision of the top-returned answer, also called top-precision (noted P0 in the following) and the recall of the system achieved by the top five answers (noted R5 in the following).

When looking for a treatment considering a question containing a pair of {pathogen; pathology} (type #1), the system's performances are presented in the Table 3. The *hybrid target* approach clearly improves the results. The system achieves a top-precision of 0.5380 for the PubMed engine and of 0.5446 for the easyIR engine. The use of a limited number of antibiotics entities avoids returning general terms, such as *Anti-Bacterial Agents*. We also evaluate the impact of using synonyms available in the controlled vocabularies we selected. The evaluations in Table 3 are reported with and without synonyms. Interestingly, the *hybrid target* approach provides several synonyms, allowing retrieving more results. Thus, *amoxycillin with clavulanate potassium* can also be mentioned as *amoxicillin-clavulanic acid* or *augmentin*.

We observe that no significant difference is reported regarding the engines, when measuring mean average precision, although it is observed that a relevance-driven ranking strategy (easyIR) returns relevant answers for more questions (coverage improved by 52%). However, the two engines, which tend to perform very similarly on average, seem not to behave similarly regarding their respective ranking power. This observation suggests that combining together the engines could be beneficial in performing the task, as proposed for instance by Fox[11]. The mean average precision (map = 0.3397), which synthesize the precision at several point of recall is maximal with the combination of the engines. Top precision, which provides a better evaluation of the task, as seen as a question-answering task, is also improved. The system achieves a P0 of 55%. The recall at five documents is of 38%, which means that out of on average three recommended antibiotics, at least one is returned in the top five answers. As for the use of keywords, they seem not to bring any significant improvement compared to the baseline system. A moderate drop of the top precision is even reported.

In almost half of the cases, the system returns a top answer not proposed by the guidelines. Indeed, different antibiotics may be used against a specific disease caused by a specific pathogen even if they are not recommended in priority. Thus, it is important to analyze errors, i.e. when the engine fails to predict the recommended antibiotic and to explore whether the answer may or may not be considered "acceptable" vs. "wrong". For this purpose we attempt to analyze the outputs regarding more generic ontological levels available in our drug reference terminology. WHO-ATC classifies drugs by anatomical, therapeutic or chemical classes. Different levels of classification are available. For instance, *cefuroxime* is a *second-generation cephalosporin*, which is a *beta-lactam antibacterial*. We aggregate antibiotics to one of their parent identifiers. Thus, our guidelines recommend *clarithromycin* to treat *gastroenteritis* caused by *campylobacter*. Using a more generic descriptor, all *macrolides* would then be considered as "acceptable" answers. For that example, the top answer returned by the system is *erythromycin*, which is also a *macrolide*. As expected, when relaxing such constraints the results showed a slight improvement, top-precision increases up to 64% (+9.91%) using PubMed and 59% (+4.18%) using easyIR. Considering higher order parents (classes or parents of parents), the top-precision rises to 77% (+23.58%) for PubMed and 81% (+26.52%) for easyIR. This means that in more than four cases out of five, the top returned antibiotic belongs to the same class as the recommended antibiotic; thus suggesting that at least a portion of the answered classified as errors, could be in fact acceptable depending on specific clinical conditions.

Search engine	easyIR		PubMed		Combination	
Measure	***P0***	***R5***	***P0***	***R5***	***P0***	***R5***
UMLS T195 target	0.1200	0.0938	0.1628	0.1481	0.1280	0.1042
WHO-ATC target	0.5103	0.3594	0.5186	0.3770	0.5402	0.3776
Hybrid target	0.5446	0.3672	0.5380	0.4172	0.5539	0.3802
Relaxing constraint (level1)	0.5864	0.4583	0.6371	0.5437	0.6109	0.4766
Relaxing constraint (level2)	0.8098	0.7682	0.7738	0.7750	0.8317	0.7024

Table 3. Evaluation of the system. P0 represents the precision of the top-answer. R5 is the recall at five documents.

The two other patterns of questions show more contrasted results. When looking for a pathology knowing a pair of {pathogen; antibiotic} (type #2), the measured top-precision is currently in the range of 13%. This is mainly due to the descriptors used, which includes extremely general categories, such as *Gastrointestinal Disease,* which will have to be filtered out to obtain cleaner and better quality answers.

Finally, the second equation, specific to surgical procedures, currently shows a top-precision of 33%, which means that in about one case out of three, the appropriate antibioprophylaxy is predicted by our rule generation approach.

Conclusion

In this report we attempted to show how text mining instruments such as question-answering engines can be used as an assistant for guidelines generation. We try to estimate how useful such a system would be when used by a domain-expert able to validate the ranked output of the guideline generator. As proof of concept, our approach was able to generate well-formed rules for up to 55% of our infectious disease benchmark. Moreover, we present here a preliminary analysis of the non-relevant answers, showing that the retrieved results are in more than 80% of the cases of the same main classes than the recommended antibiotics. Although partial, this result demonstrates that text-mining methods can automatically generate a significant subset of the expert medical knowledge, as available in a large text repository such as the MEDLINE digital library. As a future work, we plan to use alternative corpora such as ClearingHouse (http://www.guidelines.gov). Further experiments beyond antibiotic prescription are needed to establish the scalability of such approaches.

Acknowledgements

This experiment has been supported by the EU-IST-FP7 DebugIT project # 712139.

The EAGLi question-answering framework (http://eagl.unige.ch/EAGLi/) has been developed thanks to the SNF Grant # 325230-120758.

References

1. Mora Y, Avila-Agüero ML, Umaña MA, Jiménez AL, París MM, Faingezicht I. Epidemiological observations of the judicious use of antibiotics in a pediatric teaching hospital. Int J Infect Dis. 2002;6(1):74-7.
2. Iosifidis E, Antachopoulos C, Tsivitanidou M, et al. Differential correlation between rates of antimicrobial drug consumption and prevalence of antimicrobial resistance in a tertiary care hospital in Greece. Infect Control Hosp Epidemiol. 2008;29(7):615-22.
3. Gould IM. The epidemiology of antibiotic resistance. Int J Antimicrob Agents. 2008;32 Suppl 1:S2-9.
4. Harvey K, Stewart R, Hemming M, Moulds R, Use of antibiotic agents in a large teaching hospital. The impact of Antibiotic Guidelines, Med J Aust. 1983;2-217-21..
5. Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Technol Inform. 2008;136:641-6.
6. Pasche E, Teodoro D, Gobeill J, Ruch P, Lovis C. Automatic Medical Knowledge Acquisition Using Question-Answering. MIE. 2009.
7. Pestotnik SL, Classen DC, Evans RS, Burke JP. Implementing antibiotic practice guidelines through computer-assisted decision support: clinical and financial outcomes. Ann Intern Med. 1996;124(10):884-90.
8. Gobeill J, Ehrler F, Tbahriti I, Ruch P. Vocabulary-driven Passage Retrieval for Question-Answering in Genomics. TREC, National Institute of Standards and Technology. 2007.
9. Aronson A, Demner-Fushman D, Humphrey S, et al. Fusion of knowledge-intensive and statistical approaches for retrieving and

annotating textual genomics documents (2005). TREC, National Institute of Standards and Technology. 2005.
10. Ruch P, Gobeill J, Lovis C, Geissbühler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak. 2008;8 Suppl 1:S6.
11. Fox E and J Shaw J. Combination of multiple searches. *TREC*. National Institute of Standards. 1994.