Parsing Complex Microbiology Data for Secondary Use

THE OHIO STATE UNIVERSITY
WEXNER MEDICAL CENTER

Authors: Protiva Rahman[1], Albert M. Lai, PhD[2], Courtney Hebert MD, MS[1,3]
Institutions: Department of Computer Science and Engineering, Department of Biomedical Informatics, Department of Internal Medicine, The Ohio State University

Background

- Clinical microbiology culture and antibiotic susceptibility data have many secondary use applications including automated surveillance and antibiotic resistance risk prediction[1]
- Culture and antibiotic susceptibility results are often reported in complex, semi structured or free text form, making it difficult to use them for analysis. There is a need to extract relevant information from these fields and put them into structured form.
- The objective of this project is to parse these free text fields in order to identify each organism, annotate it with the appropriate unique SNOMED concept identifier (SCUI) from the UMLS, extract properties such as antibiotic susceptibility results or evidence of resistance mechanisms and associate these to the relevant organism.

Methods

Intra-abdominal cultures in patients with diagnosed intra-abdominal infection over a 5 year time period were included. To parse microbiology records, we took the following steps:

1. Tokenized the text field and removed stopwords, except for “not”;
2. Used regular expressions to extract properties (e.g., presence of resistance mechanisms).
3. If a token did not match any predefined regular expression, we determined if it matched the concept type of “bacteria, virus or yeast” in the UMLS. If it did we continued consuming tokens to find the longest string that matched. In cases where there were multiple matches, we used the Levinshtein distance to get the closest matching string (e.g., Alpha Strep ->Alpha Hemolytic Strep)
4. If an organism was not fully speciated (e.g., Escherchia/Citrobacter), we assigned the code for the common parent.
5. If the token matched the semantic type of “Antibiotic,” we look at the tokens within a distance of five to see if they match predefined words and accordingly mark them as resistant/susceptible.

Results

- We developed our algorithm on 300 cultures and tested it on the remaining 723. A subject matter expert (CH) manually validated a set of 200 parsed results, for which we report the precision, recall and F1 score.
- An organism was reported as a true positive (TP) if it was present in the culture and had the most accurate concept identifier associated, false positive (FP) if there was an identifier for an organism not present, and a false negative (FN) if an organism in the culture did not have an identifier.

	TP	FP	FN	Precision	Recall	F1
Organism	**185**	**18**	**10**	**.91**	**.95**	**.93**
Penicillinase	**190**	**0**	**10**	**1**	**.95**	**.97**
Free-text Susceptibility	**174**	**14**	**12**	**.92**	**,93**	**.93**

Conclusion

- We achieved high F1 scores using our technique of longest matching and regular expressions customized to our data.
- While we could have reduced the number of false positives by techniques such as keeping the most specific organism, and removing common parents, for most secondary purposes it is preferable to err on the side of over reporting.
- Some techniques, such as reporting the common parents on encountering a “/” character are specific to our data, however the ideas can be replicated for other free text microbiology data.
- This kind of work is vital in order to use these complex data for secondary use applications, specially for applications enabling subject matter experts to make decisions.

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the NIH under Award Number R01AI116975

Fabricated raw culture data