Assessment of Software Testing and Quality Assurance in Natural Language Processing Applications and a Linguistically Inspired Approach to Improving It

K. Bretonnel Cohen*, Lawrence E. Hunter, and Martha Palmer

Computational Bioscience Program,
University of Colorado School of Medicine,
Aurora, Colorado, USA
Department of Linguistics,
University of Colorado at Boulder,
Boulder, Colorado, USA

Abstract. Significant progress has been made in addressing the scientific challenges of biomedical text mining. However, the transition from a demonstration of scientific progress to the production of tools on which a broader community can rely requires that fundamental software engineering requirements be addressed. In this paper we characterize the state of biomedical text mining software with respect to software testing and quality assurance. Biomedical natural language processing software was chosen because it frequently specifically claims to offer production-quality services, rather than just research prototypes.

We examined twenty web sites offering a variety of text mining services. On each web site, we performed the most basic software test known to us and classified the results. Seven out of twenty web sites returned either bad results or the worst class of results in response to this simple test. We conclude that biomedical natural language processing tools require greater attention to software quality.

We suggest a linguistically motivated approach to granular evaluation of natural language processing applications, and show how it can be used to detect performance errors of several systems and to predict overall performance on specific equivalence classes of inputs.

We also assess the ability of linguistically-motivated test suites to provide good software testing, as compared to large corpora of naturally-occurring data. We measure code coverage and find that it is considerably higher when even small structured test suites are utilized than when large corpora are used.

1 Introduction

Biomedical natural language processing tools and data generated by their application are beginning to gain widespread use in biomedical research. Significant

* Corresponding author.
progress has been made recently in addressing the scientific challenges of creating computer programs that can properly handle the complexities of human language. However, the transition from a demonstration of scientific progress to the production of tools on which a broader community can depend requires that fundamental software engineering requirements be addressed. Software for medical devices has the benefit of explicit quality assurance requirements per Section 201(h) of the Federal Food, Drug, and Cosmetic Act; Title 21 of the Code of Federal Regulations Part 820; and 61 Federal Register 52602 [8] (p. 7). However, unless it is embedded in a medical device, biomedical natural language processing software is not currently subject to federal quality assurance requirements.

This paper represents the first attempt to characterize the state of one portion of the diverse world of computational bioscience software, specifically biomedical natural language processing applications, with respect to software testing and quality assurance. We assay a broad range of biomedical natural language processing services that are made available via web sites for evidence of quality assurance processes. Our findings suggest that at the current time, software testing and quality assurance are lacking in the community that produces biomedical natural language processing tools. For the tool consumer, this finding should come as a note of caution.

2 Approach to Assessing the State of Natural Language Processing Applications with Respect to Software Testing and Quality Assurance

We looked at twenty web sites offering a variety of text-mining-related services. In the body of this work, we never identify them by name: following the tradition in natural language processing, we do not want to punish people for making their work freely available. Our purpose is not to point fingers—indeed, one of our own services is every bit as lacking in most or all of the measures that we describe below as any. Rather, our goal is to allow the community to make a realistic assessment of the state of the art with respect to software testing and quality assurance for biomedical natural language processing systems, with the hope of stimulating a healthy change.

The claim to have produced a useful tool is a commonplace in the biomedical natural language processing literature. The explicitly stated motivation for much work in the field is to assist in the understanding of disease or of life, not to advance the state of computer science or of understanding of natural (i.e., human) language. (In this, the biomedical natural language processing community differs from the mainstream NLP community, which at least in theory is motivated by a desire to investigate hypotheses about NLP or about natural language, not to produce tools.) Software is widely known to be characterized by “bugs,” or undesired behaviors—[15] reviews a wide range of studies that suggest an industry average of error rates of 1 to 25 bugs per thousand lines of