NL+SE 2015:

NSF Interdisciplinary Workshop on
Statistical NLP and Software Engineering
Oct 25-27, 2015 at Microsoft Research, Redmond, WA
(by invitation only)

Statistical Natural Language Processing (NLP) techniques and tools have progressed rapidly over the past decade, leading to great advances in areas such as parsing and machine translation. Much of this progress was driven by algorithms that exploit big data and big compute resources. Looking toward Software Engineering (SE), there are many promising opportunities for applying NLP techniques to formal languages. Recently, investigations have begun to understand to what extent large code corpora are amenable to analysis using statistical NLP models and algorithms, so that the revolutionary advances in speech recognition, translation, and natural language understanding can be applied in SE. Meanwhile, the NLP community has begun addressing problems with formal representations as their target, such as semantic parsing and end-user programming.

The workshop will bring together an international group of researchers in Statistical NLP, Programming Languages, Software Engineering and related fields for an intensive period of discussion. The specific aims are:

- To help NLP researchers understand the range of possible applications of statistical methods in software development, and the wealth and diversity of available code corpora.
- To help SE researchers understand the capabilities and limitations of large-corpus, statistical methods over textual data.
- To outline an agenda for future work in the area.
- To help bring about new interdisciplinary collaborations.

Topics for discussion include but are by no means limited to the following:

- Statistical language models as applied to programming: code prediction, code correction, and standards discovery, enforcement, etc.
- Statistical language translation between formal languages: semi-automated software porting, code transformation, code styling, and refactoring, etc.
- Methods for inferring program meaning: de-obfuscation, annotating types, segmentation, requirements tracing, etc.
- Cross-lingual methods applied to pairs of formal and natural languages: code retrieval, code summarization, inconsistency detection, assertion generation/suggestion, etc.
- Bi-lingual, cross-language methods for code retrieval, code summarization, inconsistency finding, assertion generation/suggestion, etc.
- End-user programming: code synthesis from natural language, especially in domain specific languages
- Knowledge Acquisition from Code Corpora: Idiom Mining, Ontologies, Fact acquisition.

Agenda: The first day, Sunday, will be focused on tutorials: statistical NLP tutorials will introduce language modeling, classification, and structure prediction techniques to software engineers, software engineering tutorials will introduce promising corpora, current techniques, and promising future directions. The 2nd and 3rd days will be the workshop proper, with a combination of presentations and working groups. A final report summarizing the ideas and new directions discussed at the workshop submitted after workshop completion.

Submission Call: We ask that invitees submit a 1-2 page statement of interest, stating their commitment to attend the workshop, and outlining their research skills and their interests in the area. The paper is primarily aimed at helping us organize the workshop program. Please send us a Accept/Decline email with a single sentence on your proposed topic in one week, and a paper by Sept 1, 2015.

Sponsorship: We anticipate funding from NSF and other sources. NSF funds (pending) will be used to cover reasonable (economy-class) travel and accommodations for most *academic* invitees (both US-based and foreign). Priority will be given to accommodating attendees from under-represented groups, as well as individuals who are not otherwise independently funded to attend this workshop. Microsoft Research will provide the venue as well as some meals and refreshments.

Organizers: Program chairs: Chris Quirk (Microsoft), Prem Devanbu (UC Davis), Dana Movshovitz-Attias (CMU)

Steering Committee:
Charles Sutton (Edinburgh)

Daniel Tarlow (Microsoft Research, Cambridge)

Dawn Lawrie (Loyola)

Dennis Poshyvanyk (Willam & Mary)

Ray Mooney (University of Texas Austin)

Tao Xie (University of Illinois)

William Cohen (Carnegie Mellon University)

NSF sponsors: Sol Greenspan (NSF) and Tatiana Korelsky (NSF)