Engaging high school students in interdisciplinary studies through the Computational Linguistics Olympiad

Dragomir Radev, University of Michigan
radev@umich.edu

Lori Levin, Carnegie Mellon University
lsl@cs.cmu.edu

Introduction

The USA and Canada have been facing a reduction in enrollments in computer science courses and a drop in the number of offerings of high school courses in computing and related subjects. In this report, we will discuss a recent attempt to reinvigorate the stream of high school students interested in this topic. We hope that more students will become interested in computer science if they can pursue interesting applications than if they are only learning to program for its own sake. With this goal in mind, the North American Computational Linguistics Olympiad (NACLO, http://www.cs.cmu.edu) has been held in the United States and Canada since 2007. Its purpose is to attract high school students to studies and careers that involve linguistics, computation, and human language technologies. Problems are self-contained and can be solved without special, pre-existing knowledge. The contest is targeted at high school students, but middle school students are also invited to participate.

Problem Design

NACLO includes problems in both traditional linguistics and computational linguistics.

The traditional problems are in the style of the International Linguistics Olympiad (ILO) and include deciphering texts in lesser known languages (such as Aymara from Bolivia, Hmong from Cambodia, and Huishu from India), as well as number, kinship, and calendar systems. We have
collaborated closely with colleagues in linguistics departments to collect original problems in this genre.

The *computational linguistics problems* are the most innovative component of the contest. Since both of us are computational linguists, we have been able to engage a number of colleagues in suggesting problems as well as drawing from the literature. So far, we have included problems in parsing, optical character recognition, text summarization, information answering, spelling correction, lexical acquisition, speech processing, and finite state automata.

All 33 problems used in the three editions of NACLO have been entirely self-contained. Since linguistics and language technologies are not taught in high schools (unlike the subjects of the other five major international Olympiads in Mathematics, Physics, Chemistry, Informatics, and Biology), we could not and did not expect students to have any prior preparation. Instead, our problems, while requiring some intuition about language (in general; not any particular language), can be solved by logical and algorithmic thinking alone. The specific skills needed are, for both the traditional and computational problems, search space formulation, search, data abstraction, dealing with incomplete and contradictory evidence, generalization, and so on.

**Participation in NACLO**

The contest consists of two rounds each year: an open round and an invitational round. Table 1 shows the number of participants in each of the last three years. The participants in NACLO have come from almost 30 states and several Canadian provinces. We estimate that about 49% of the participants are female. The top participants have come in roughly equal numbers from public schools and from private schools.

One important early decision that contributed to the popularity of NACLO was to hold the contest in a distributed fashion at a large number of high school sites and university sites. University sites provide a place for students from many high schools to come together and meet other students with similar interests. University sites may also offer demos and presentations on the day of the contest. At the same time, allowing high school sites makes it possible for any interested student, no matter how far from a university, to take part in the contest.
In the next two years we plan to expand the contest to include automatically gradable problems as well as problems that require a computing environment. Also, since production of problems is labor intensive, we are collaborating with other English speaking countries, such as Australia, Ireland, India, and Great Britain to produce and share a larger collection of problems in English. Each of these countries will participate separately in the International Linguistics Olympiad (see below).

**Outreach to High Schools**

NACLO is publicized in various ways in the cities in which it is held, usually by direct contact with high schools or through newspaper articles. Faculty who host NACLO at their university usually visit local high schools to provide training sessions and register students for the contest. When publicizing NACLO in high schools we have been focusing on certain aspects of linguistics and computer science. With respect to linguistics, we emphasize that languages have rules and patterns that native speakers may not be aware of; that there are procedures by which these rules and patterns can be discovered in one’s own language; and that the same procedures can be used to discover rules and patterns in languages other than one’s own. We use computational linguistics as a way to emphasize that computer science is not just about machines or code, but also about how to structure and solve a problem. We then introduce some challenging problems in language technologies such as web search, telephone dialogue systems, speech recognition, and machine translation.

**Sponsorship**

NACLO has been primarily sponsored by the National Science Foundation and other government agencies as well as companies such as Google and Cambridge University Press, local sponsors, the North American Chapter of the Association for Computational Linguistics, and individual donors. Our universities, Carnegie Mellon University (for Levin) and the University of Michigan (for Radev), have also strongly supported our involvement with the contest.

**The International Linguistics Olympiad**
The eight highest scoring students in NACLO each year have been invited to be part of the US teams that participated in the international Linguistic Olympiads (ILOs) held in Russia (2007), Bulgaria (2008), and Poland (2009). As of the time of writing this material, Sweden, the USA, and Slovenia are likely to host the international contest in the near future. The ILOs have been in existence since 2003 but US teams started taking part in them only in 2007.

In 2007, the US team won a gold medal for the highest score at the individual contest (Adam Hesterberg) and one of the US teams tied with one of the teams from Russia for first place at the team contest (Josh Falk; Rebecca Jacobs; Michael Gottlieb; and Anna Tchetchetkine). In 2008, the US team was even more successful, bringing home one gold medal (Hanzhi Zhu), two silver medals (Morris Alper and Anand Natarajan), and three bronze medals (Rebecca Jacobs, Guy Tabachnick, and Jeffrey Lim) as well as a first place (tied) in the team contest (Morris Alper, Rebecca Jacobs, Jae-kyu Lee, and Hanzhi Zhu). In 2009, the US team obtained for the third time in a row (and for the first time, without tying another team) the team gold (Rebecca Jacobs, Anand Natarajan, Alan Huang, and Morris Alper) as well as one individual silver (Rebecca Jacobs) and three bronzes (John Berman, Sergei Bernstein, and Alan Huang).

Out of a total of 18 team members over the last three years, three have gone or are going to Princeton and three to MIT. The University of Chicago and Stanford are getting two each and the rest of the graduating students will be attending CalTech, Cornell, Harvard, and the University of Washington. The other four have not graduated yet. Most of these students are majoring in Mathematics, Computer Science, Languages, and Linguistics, or some combination of these fields.

The web site for NACLO is www.naclo.cs.cmu.edu. The site includes information about recent NACLOs, about 300 sample problems, as well as information about starting clubs in high schools, about future participation, and about hosting new sites.
Table 1: Statistics about the first three years of NACLO

<table>
<thead>
<tr>
<th>Year</th>
<th>Open round participants</th>
<th>Invitational round participants</th>
<th>Number of sites</th>
<th>Performance at the ILO</th>
<th>Best individual student at the ILO</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>195</td>
<td>n/a</td>
<td>3 univs, 20 schools</td>
<td>4 prizes</td>
<td>Adam Hesterberg from Garfield HS, Seattle, WA</td>
</tr>
<tr>
<td>2009</td>
<td>1080</td>
<td>135</td>
<td>27 univs, 65 schools</td>
<td>7 prizes</td>
<td>Rebecca Jacobs from Harvard-Westlake HS, Los Angeles, CA</td>
</tr>
</tbody>
</table>