

**NUS Graduate School for Integrative Sciences and Engineering
Research Project Write-up**

Title of Project : Machine Learning and Human Computation Synergy for Machine Translation

Name of Supervisor : KAN Min-Yen

Contact Details: kanmy@comp.nus.edu.sg

Short Description

In recent years, the notion of human computing has been introduced to complement machine learning systems. The idea is to harness what people are good at but machines are poor at. In this way we can 1) create Turing tests to separate people from 'bots and 2) create annotated data for training and improving research systems. This approach is also called human-based computation. It is a technique leverages differences in abilities and alternative costs between humans and computer agents to achieve human-computer interaction [14]. Many tasks are trivial for humans but continue to challenge even the most sophisticated computer programs [9][10]. Thus, the intelligent combination between computer and human to solve complex tasks becomes a promising approach.

Luis von Ahn et al. [9][10] proposed a new definition relating to human-based computation which is called game with a purpose (GWAP). GWAPs are games in which limitations by computer computation will be solved by employing human guidance. The data generated by effect of game plays an important role to solve computational problems and trains AI algorithms. Such games are simultaneously made for funny entertainment of human and obtaining useful computational purposes for computer. The main purpose of these games is to collect data as much as possible with the guidance of human and intelligent support of computer.

So far, while successful, the work has not showed how humans can improve from this relationship. Thus far, the idea of human computing has been just to formulate GWAPs, for entertainment. However, such an interaction opportunity also makes it an ideal situation to train or tutor the game player - in the guise of Game Based Learning.

In this project, the student will study and implement this framework for acquiring machine translation training data. The goal is to implement a system to train second language learners of English and Chinese, while gathering useful training data for machine translation systems. By gathering useful training data, we mean that the data annotated by the humans must be helpful in improving eventual translation quality. By training second language learners, we mean that the users are measurably improved at their language production or consumption.

The student taking on this project must have a solid background in statistical MT and ML. In investigating this project, he/she must be independent and interdisciplinary, bringing in and relating work from HCI/UI, language learning, computer based

training communities as well.

References - MT & SMT

- [1]. Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. "The mathematics of statistical machine translation: parameter estimation". *Computational Linguistics*, 1993, 19(2):263-311.
- [2]. Philipp Koehn, Franz Josef Och, and Daniel Marcu. "Statistical phrase-based translation". In *Proc. of the HLT-NAACL 2003 conference*, pages 127-133, Edmonton, Alberta, Canada, May 2003.

Human Computing

- [9]. Luis von Ahn and Laura Dabbish. "General Techniques for Designing Games with a Purpose". *Communications of the ACM*, August 2008. Pages 58-67.
- [10]. Luis von Ahn. "Games with a Purpose". *IEEE Computer Magazine*, June 2006. Pages 96-98.
- [11]. Von Ahn, L., Kedia, M., and Blum, M. "Verbosity: A Game for Collecting Common-Sense Knowledge".
- [12]. Luis von Ahn and Laura Dabbish. "Labeling Images with a Computer Game".
- [13]. Von Ahn, L., Liu, R., and Blum, M. "Peekaboom: A Game for locating objects in images".
- [14]. http://en.wikipedia.org/wiki/Human-based_computation
- [15]. http://en.wikipedia.org/wiki/Turing_test
- [16]. Marc Prensky, *Digital Game Based Learning*

(English) Language Learning

- [17]. Lian Pek, "Mad About English" 2008, Documentary Film.