COMPUTER PREPARES TO READ LIKE HUMANS

*The Never-Ending Language system runs 24*7 to extract facts from text found in web pages to improve its reading competence*

LESLIE D'MONTE

Can computers learn to read? A Carneige Melon University research team, that includes an Indian PhD student believes so.

The research project attempts to create a computer system that learns over time to read the web.

For the last ten months, the computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day.

First, the system attempts to "read", or extract facts from text found in hundreds of millions of web pages [for instance, playsInstrument(George_Harrison, guitar)] and then, it attempts to improve its reading competence, so that it can extract more facts from the web, more accurately on the following day.

NELL runs 24*7 to perform two ongoing tasks. It has currently acquired a knowledge base of nearly 440,000 "beliefs" that it has read from various web pages.

'Read the Web', as it is called, aims at eventually building a never-ending language learner (hence, NELL) — a computer agent that runs forever and extracts, or reads, information from the web daily to populate a growing structured knowledge base. Morever, it must learn to perform a specific task better than what it had achieved on the previous day.

NELL has currently acquired a knowledge base of nearly 440,000 'beliefs'

For the first six months, NELL was allowed to run without human supervision, learning to extract instances of a few hundred categories and relations, resulting in a knowledge base comprising approximately a third of a million extracted instances of these categories and relations.

The inputs to NELL include an initial ontology defining hundreds of categories (for example, person, sportsTeam, fruit, emotion) and relations [like,playsOnTeam(athlete,sportsTeam), playsInstrument(musician,instrument)] that NELL is expected to read about, and 10-15 seed examples of each category and relation. Given these inputs, plus a collection of 500 million web pages and access to remainder of the web through search engine application programming interfaces (APIs).

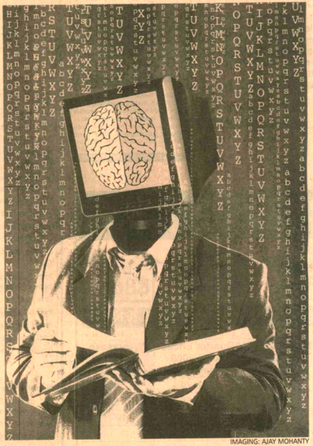
NELL extracts new instances of categories and relations. In other words, it finds noun phrases that represent new examples of the input categories (for example, "Barack Obama" is a person and politician), and finds pairs of noun phrases that correspond to instances of the input relations (for example, the pair "Jason Giambi" and "Yankees" is an instance of the playsOnTeam relation). These new instances are added to the growing knowledge base of "beliefs".

NELL uses a variety of methods to extract "beliefs" from the web. Much of its current success is due to its algorithm for coupling the simultaneoůs training of many extraction methods. In July, a spot test showed the average precision of the knowledge base was approximately 87 per cent across all categories and relations.

NELL, however, makes many mistakes too. For instance, for the category bakedGood, it learns the pattern "X are enabled in" because of the believed instance "cookies." This leads it to extract "persistent cookies" as a candidate bakedGood. The probability for phrases that end in "cookies" is high and so "persistent cookies" is promoted as a believed instance of bakedGood.

When it comes to card games, the cardGame category seems to suffer from the abundance of web spam related to casino and card games, which results in parsing errors and other problems. As a result of this noise, NELL ends up extracting strings of adjectives and nouns like "deposit casino bonuses free online list" as incorrect.

The computer also finds it difficult to associate product names with more general nouns that are somehow related to the product but do not correctly indicate what type the product is, (for example,"Microsoft Office", "PC").



IMAGING: AJAY MOHANTY

The research team is still trying to understand what causes it to become increasingly competent at reading some types of information, but less accurate over time for others. "It is not perfect, but NELL is learning..," said the research team on the project website.

The team includes professors Tom Mitchell and William Coehn and an Indian PhD Student (Language Technologies Institute) Jayant Krishnamurthy.

The computer also makes use of Yahoo!'s M45 computing cluster to efficiently extract statistics from the half billion web pages. Financial support for the research has been provided in part by DARPA, the National Science Foundation (NSF), Google, and the Brazilian agency CNPq.

Incidentally, scientists at universities, government labs, and technology companies like Google, Microsoft and IBM have similar pursuits. While the online search giant has 'Google Squared, IBM's project is code-named 'Watson' after its founder Thomas J Watson. The IBM computing system, unveiled last year, is designed to rival the human mind's ability to understand the actual meaning behind words, distinguish between relevant and irrelevant content, and ultimately, demonstrate confidence to deliver precise final answers. Watson will not be connected to the Internet, or have any other outside assistance.

(The author, on a sabbatical from Business Standard, *is an MIT Knight Science Journalism Research Fellow 2010-11)*