NELL: The Never-Ending Language Learning System[1]

Estevam R. Hruschka Jr.[2]

Federal University of Sao Carlos, Sao Carls SP – Brazil

e-mail: estevam@dc.ufscar.br

ABSTRACT

Never-Ending Language Learner (NELL)[1] is a computer system that runs 24/7, forever, learning to read the web. extract (read) more facts from the web, and integrate these into its growing knowledge base of beliefs; and ii) learn to read better than yesterday, enabling it to go back to the text it read yesterday, and today extract more facts, more accurately. This system has been running 24 hours/day for over four years now. The result so far is a collection of 70 million interconnected beliefs (e.g., servedWith(coffee, applePie), isA(applePie, bakedGood)), that NELL is considering at different levels of confidence, along with hundreds of thousands of learned phrasings, morphological features, and web page structures that NELL uses to extract beliefs from the web .

1 INTRODUCTION

Despite tremendous progress in machine learning over the past decades, we still have very limited number of attempts to build machine learning systems that learn cummulatively forever, using what they learned yesterday to improve their ability to learn tomorrow, and improving indefinitely. We seek to build such a system, in the domain of natural language understanding (reading the web). We call this system NELL (Never-Ending Language Learner).

The main motivation for building NELL is based on the belief that we will never really understand machine learning until we can build machines that, like people, have the following characteristics: i) *learn many different types of knowledge or functions;* ii) *from years of diverse, primarily self-supervised experience;* iii) *in a staged curricular fashion, where previously learned knowledge enables learning further types of knowledge;* iv) *where self-reflection and the ability to formulate new rep- resentations and new learning tasks enable the learner to avoid stagnation and performance plateaus*

The domain choice (*reading the web*) is mainly motivated because we believe AI community will not be able produced *Natural Language Understanding* systems (which should go beyond *Natural Language Processing*) until we have computer systems that react to arbitrary sentences by saying one of the following: *I understand, and I already knew that;* or, *I understand, and I didn't know, but I accept that;* or *I understand, and I disagree because ...*

The input to NELL is an initial ontology, with categories and relations, as well as instances (labeled training examples) of these categories and relations. Given this input, our goal is for NELL to run 24 hours/day, 7 days/week, forever, interacting with human trainers for up to an hour per day. On each day, NELL must accomplish two things:

1. **Performance task**: Each day it must extract more factual beliefs from the web in order to further populate its knowledge base, according to the given ontology.
2. **Learning task**: Each day it must learn to read better than it could the previous day.

NELL is evaluated by its success in achieving both of these tasks. To evaluate its success at the first task, we evaluate the correctness and breadth of the beliefs it extracts. To evaluate its success at the second task, we measure the change over time in its competence at task one. For example, we can send it to a sample of the same web sources it visited yesterday, and measure whether it extracts more facts more accurately today than it did yesterday.

2 NELL KEY FEATURES AND CURRENT STATE

Based on the main characteristics mentioned in the previous section, NELL has been built and is running since January, 2010. Currently the system follows its Cumulative, Staged Learning showing that performing a learning task *X* can help improving the ability to learn a different task *Y*. NELL's current ability to learn includes: i) classify noun phrases (NPs) by category; ii) classify NP pairs by relation; iii) discover rules and patterns to predict new relation instances; iv) learn which NPs (co)refer to which latent concepts; v) discover new relations to extend the initial ontology; vi) learn to assign temporal scope to beliefs; vii) learn to *microread* single sentences; viii) vision: co-train text and visual object recognition; ix) goal-driven reading: predict, then read to corroborate/correct. In addition we have started working on making NELL a conversational agent on Twitter and we plan to add a robot body to NELL.