

NELL

NEVER-ENDING LANGUAGE LEARNING SYSTEM

JELENA JOVANOVIĆ

jeljov@gmail.com ; <http://jelenajovanovic.net>

Read the Web @ CMU

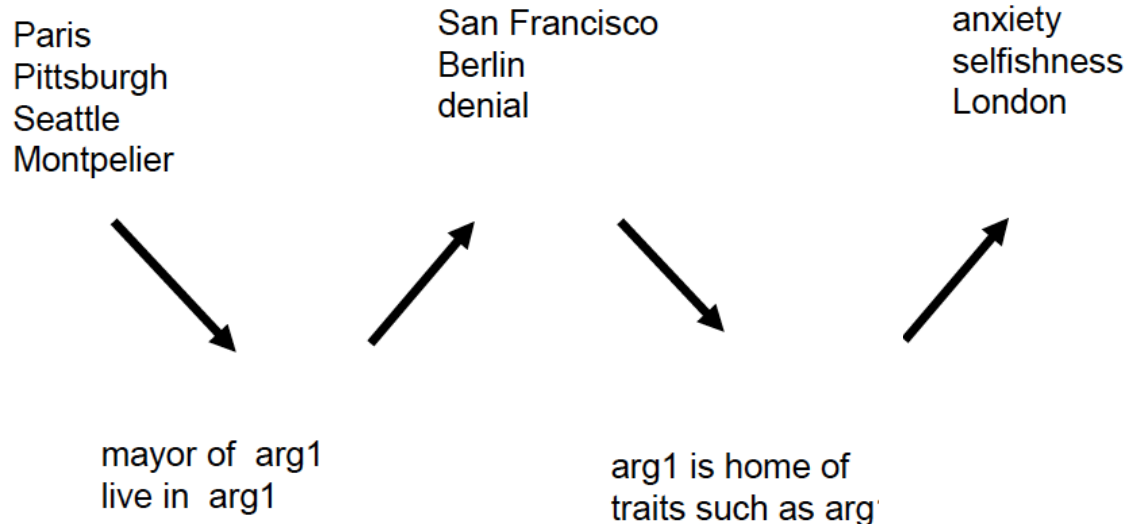
- A research project at the Carnegie Mellon University
 - <http://rtw.ml.cmu.edu/rtw/>
- Research Goals:
 - Develop a never-ending machine learning system for extracting structured information from unstructured Web pages
 - The development of the world largest structured KB that
 - reflects the factual content of the Web,
 - continually grows in terms of both predicates and instances,
 - could be useful to many AI efforts

Read the Web @ CMU (2)

The underlying assumptions

- the vast redundancy of information on the Web - many facts are stated multiple times in different ways
- different learning methods can be used to extract complementary information, and to validate the extracted information items
- periodic human feedback can prevent the ‘semantic drift’

An example of ‘semantic drift’



Read the Web @ CMU (3)

- Input for learning:
 - Ontology defining hundreds of categories of things and semantic relations between those things
 - Categories: person, athlete, sportsTeam, fruit, emotion,...
 - Relations: playsOnTeam(athlete,sportsTeam),...
 - Seed: 10-15 examples for each category and relation
 - Access to a huge collection of Web pages
 - >500M Web pages from the [ClueWeb09 dataset](#)
 - Search engine APIs

Never Ending Language Learner (NELL)

NELL is an implementation of the Read the Web approach























It performs 2 tasks each day, 7 days per week:

- *Reading task*: extract new instances of categories and relations from texts on the Web, and thus extend the KB
- *Learning task*: learn to ‘read’ better each day, as evidenced by the ability to extract more information more accurately
 - the learning components continuously retrain themselves using the growing KB as a set of training examples


One can follow NELL while it ‘reads’,
and help it *learn* to ‘read’ better

Recently-Learned Facts

Refresh

Instance	Iteration	date learned	confidence		
<u>chris briton</u> is a <u>Mexican person</u>	818	03-mar-2014	90.4		
<u>oscar villarreal</u> is an <u>athlete</u>	818	03-mar-2014	100.0		
<u>spinal gray matter</u> is a kind of <u>brain tissue</u>	821	11-mar-2014	90.9		
<u>home goods</u> is a <u>retail store</u>	818	03-mar-2014	90.7		
<u>mt kilimanjaro</u> is a <u>mountain range</u>	820	08-mar-2014	91.7		
<u>los angeles</u> is the <u>home city of the sports team southwestern university</u>	823	19-mar-2014	93.8		
<u>vladimir guerrero</u> is a sports coach <u>also known as ervin santana</u>	820	08-mar-2014	96.3		
<u>commerzbank</u> is a bank that <u>bought dresdner bank</u>	823	19-mar-2014	100.0		
<u>new york mets</u> is a sports team that <u>won the pennant</u>	821	11-mar-2014	98.4		
<u>oswald</u> and <u>john f kennedy</u> are <u>siblings</u>	821	11-mar-2014	93.8		



NELL on Twitter: @cmunell




NELL
@cmunell

I am a machine reading research project at Carnegie Mellon, periodically tweeting facts I read. Please follow me, and reply with corrections so I can improve!

Pittsburgh PA · rtw.ml.cmu.edu

TWEETS	FOLLOWING	FOLLOWERS	
18.9K	644	2,334	  Follow


Tweets



NELL @cmunell · 50m

True or False? "Josh Schwartz" is an [#Actor](http://bit.ly/1kfjKJK) (bit.ly/1kfjKJK)


Expand ↩ Reply ↻ Retweet ★ Favorite ... More



NELL @cmunell · 2h

True or False? "opensecrets.org" is a [#MachineLearningConference](http://bit.ly/1mNnC8o) (bit.ly/1mNnC8o)


Expand ↩ Reply ↻ Retweet ★ Favorite ... More



NELL @cmunell · 3h

True or False? "Rosina Wachtmeister" is an [#Architect](http://bit.ly/1IR1sIX) (bit.ly/1IR1sIX)

Expand ↩ Reply ↻ Retweet ★ Favorite ... More



NELL @cmunell · 5h

True or False? "Stupid Rock" is a [#GenreOfMusic](http://bit.ly/1pAKaX) (bit.ly/1pAKaX)

Expand ↩ Reply ↻ Retweet ★ Favorite ... More

NELL (4)

- Combined use of different learning components:
 - A free-text extractor which (interchangeably) learns and makes use of contextual patterns to extract instances of categories and relations
 - patterns like “mayor of X” and “X plays for Y ”
 - A component that extracts novel instances from semi-structured Web data (e.g., tables, lists)
 - A set of binary (logistic regression) classifiers - one per category - which classify NPs based on various morphological features
 - A component that learns probabilistic rules for inferring new kinds of relations from already learned relations

NELL (5)

- Presently, NELL is *not* a targeted reader
 - it picks facts from here and there, without domain/topic specific focus
- An announcement for this year: “knowledge on demand”
 - one would give NELL a query, and NELL would do targeted reading to be able to answer the given query
 - this would enable both semantic labeling of Web content, and construction of domain-specific structured KBs