



Seminar Information Extraction

Never Ending Language Learning

Lukas Veneziano

Motivation

- Knowledge base with internet facts
- Grows every day/iteration
- Gets every day more accurate
- Uses state of the art techniques of IE
- NELL started in the Year 2010 and is still computing

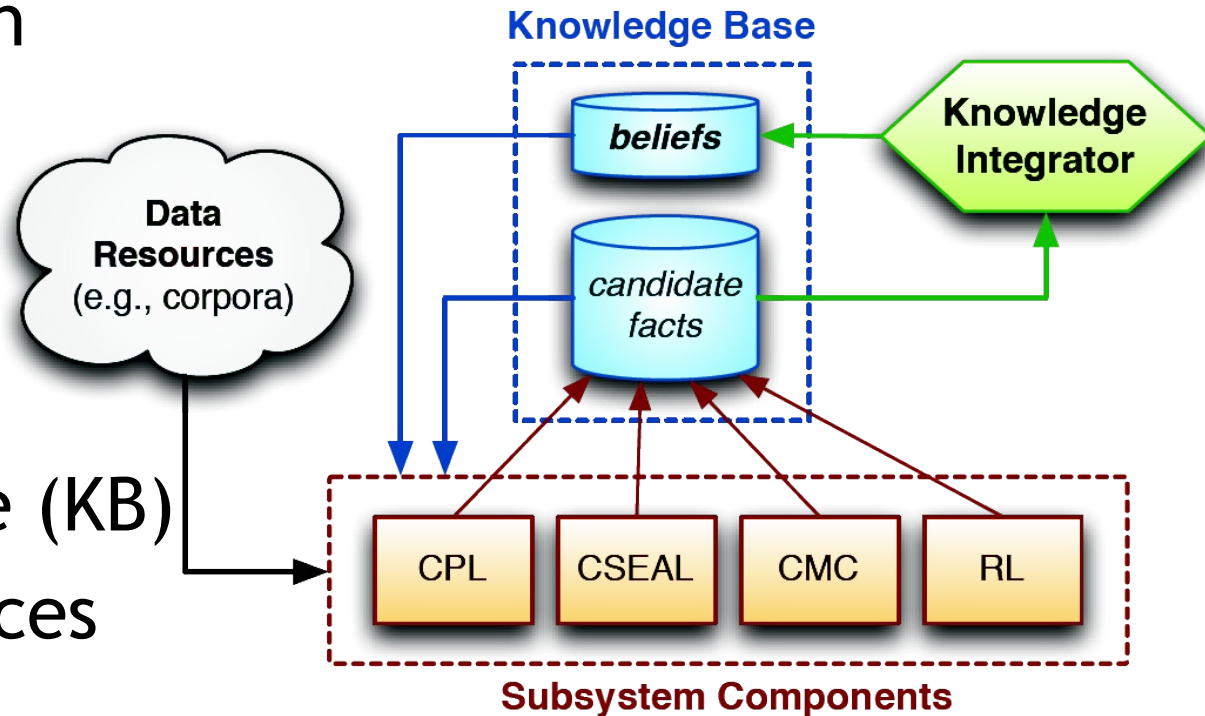
How does it work?

NELL...

- ... crawls the net for information
- ... evaluates the information
- ... stores beliefs in a knowledge base
- ... iterates over time with the received knowledge to get better results
- ... graduates (or degrades) old beliefs after one iteration

How does it work?

- Subsystems with different approaches
- Knowledge integrator (KI)
- Knowledge base (KB)
- External resources



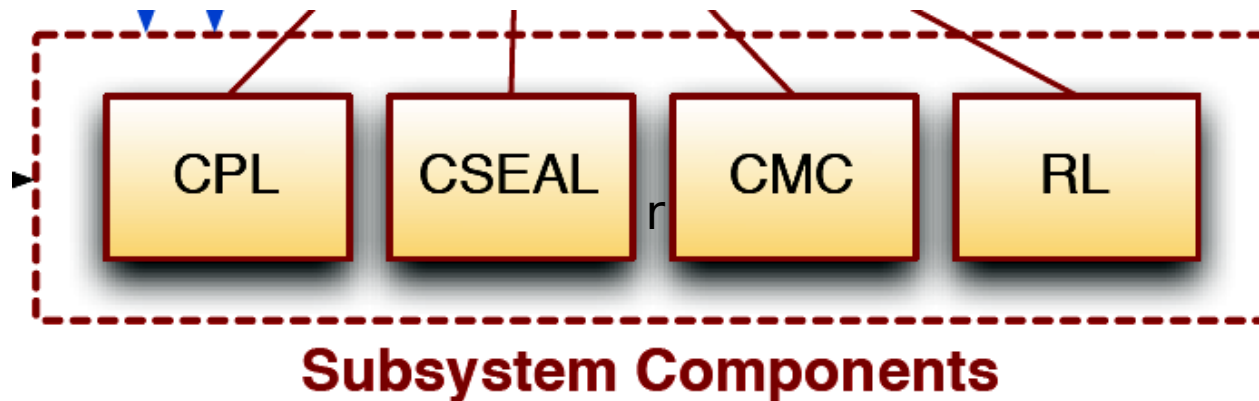
Knowledge Integrator



- Decides if a candidate gets to a belief
- Candidates get propose
 - Probability
 - Summary

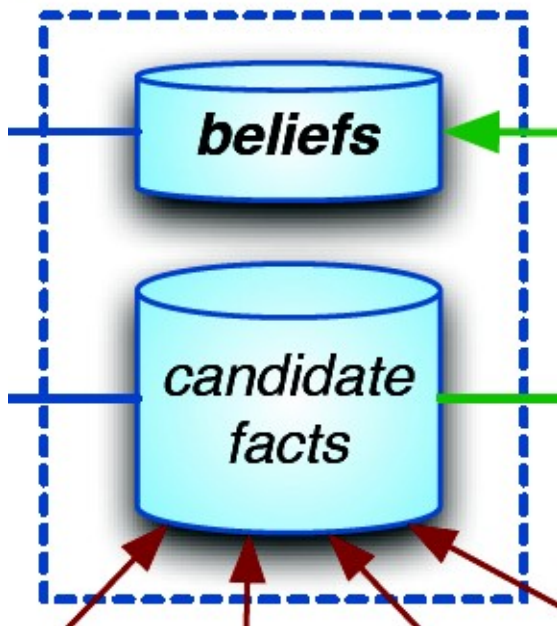
Subsystems

- Read from data resources and KB
- Propose to KI
- Uncorrelated errors
- Multiple types of inter-related knowledge



Knowledge Base

Knowledge Base



- Beliefs
(posterior > 0.9)
- Candidate facts
(posterior ≤ 0.9)
- Retains source justification
for each belief

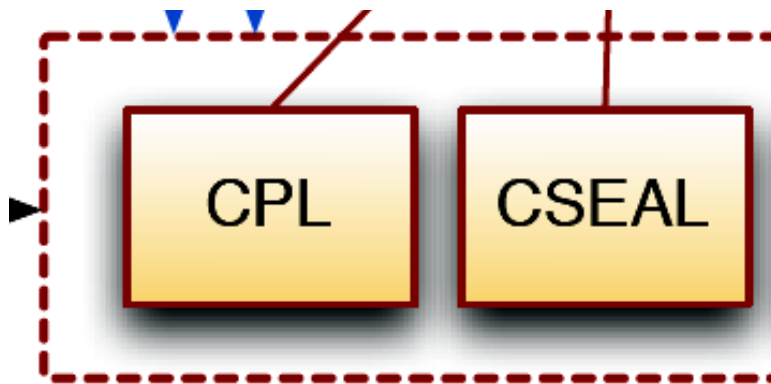
Data Resources



- Corpora
- Internet pages
 - in the original training set:
500 mio. web pages
- etc.

Two examples for subsystems

- I. Coupled Pattern Learner
- II. Coupled SEAL



I. Coupled Pattern Learner

- Free text extractor which learns contextual patterns and relations
- Uses part-of-speech tagging
- Gives probabilities $[1, 0.5]$
- Relative to number of promoted patterns c
- Formula: $1 - 0.5^c$
- In experiments openNLP was used to check accuracy

I. Coupled Pattern Learner

Pattern learning:

- extracts category patterns from promoted categories if *a noun is followed by a verb*
- extracts relation patterns if *arguments of a promoted relation occur more than once*

Instance learning:

- extracts relation instances *if both place-holders are valid for their categories*
- Extracts category instances *if proper noun phrase with valid specification is found*

II. Coupled SEAL

- Semi-structured extractor, queries the internet with beliefs
- Used to get new instances of a predicate
- Uses mutual exclusion relationships
- Same method for probability as CPL

More information in [3]

II. Coupled SEAL

- Searches web-pages with some predefined seeds for new elements
- Original SEAL had no concept of mutual exclusion
 - CSEAL first extracts new elements by original SEAL Algorithm and then filters for candidates that violate coupling
 - If carlsOfBrand(BMW) then not carlsOfBrand(Audi)

Demo

Many improvements

- Over time there were published over 20 new papers from the group with improvements for NELL
 - One talks about human supervision in a NELL-system-implementation

Human Supervision

- Two approaches:
 - Active and passive looking for human collaboration
- Passive approach on the web-site
- Active approach autocratically seeks for answers in different communities
- In NELL both approaches are connected

Problems of Human Supervision

- Computer must autonomously get help from humans
- Which knowledge should be out to humans attention?
- Who are the humans that the machine should look for help?
- How to understand humans answers?
- How to infer from humans answers?

Conversing Learning

- NELL uses CL to actively seeking for information from human users
- only few facts are going to be validated by community
- Advancement to the SS-Cloud algorithm

Tasks:

- I. Take facts from the KB
- II. Build a human understandable question
- III. Query the web community with the question
- IV. Gather and resolve the answer (classify as positive or negative)
- V. Combine the answers and produce a combined opinion

Twitter: @cmunell

- <https://twitter.com/cmunell>

Twitter vs. Yahoo!Answers

Unresolved Answers:

- Twitter:

5,5%

	Approved	Rejected	Unresolved
Twitter	51	17	4
Yahoo	124	168	58

- Yahoo!Answers:*

16,5%

- Twitter and Yahoo!Answer users disagree:*

45%

- Unfortunately there is no data for improvements on NELLs accuracy

- But earlier human supervision sessions improved from 71% to 87% accuracy

Results of the first NELL

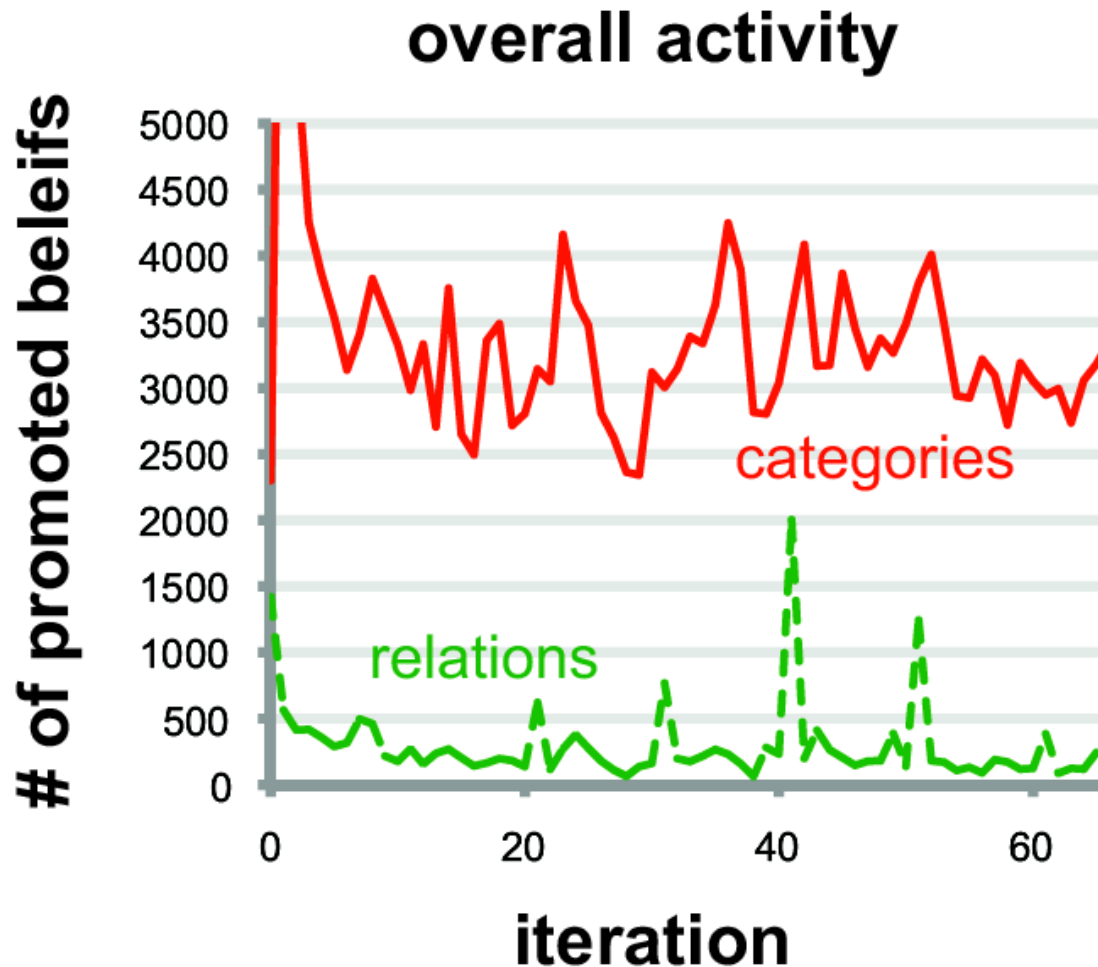
Testconditions:

- Every iteration: CPL, CSEAL, CMC
- Every 10 iterations: RL
 - Proposed outputs were filtered by a human
(few minutes per iteration)
- Training Set:
 - 123 categories with each 10-15 seed instances and 5 seed patterns
 - 55 relations with each 10-15 seed instances and 5 negative instances
 - (some?) relationships between categories

Results of the first NELL

- Runtime of 67 days
 - 66 iterations
- 242 453 new beliefs
 - 95% instances of categories
 - 5% relations

Results of the first NELL



Results of the first NELL

Iterations	Estimated precision (%)	#promotions
1-22	90	88,502
23-44	71	77,835
45-66	57	76,116

Questions?

Thank you for your attention.

- [1] **Toward an Architecture for Never-Ending Language Learning**, Andrew Carlson, Justin Betterige, Brian Kisiel, Bur Settles, Estevam R. Hruschka Jr. And Tom M. Mitchel, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, USA
- **Conversing Learning: active learning and active social interaction for human supervision in never-ending learning systems**, Saulo D. S. Pedro and Estevam R. Hruschka Jr., Federal University of Sao Carlos, UFSCar Sao Carlos, Brazil
 - **Coupled Semi-Supervised Learning for Information Extraction**, Andrew Carlson, Justin Betterige, Richard C. Wang, Estevam R. Hruschka Jr. And Tom M. Mitchel, Federal University of Sao Carlos, UFSCar Sao Carlos, Brazil, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, USA
 - <http://rtw.ml.cmu.edu/rtw/overview>, viewed, 2013-12-10, 23:26
 -