NELL2RDF: Reading the Web, and Publishing it as Linked Data

José M. Giménez-García[1], Maísa Duarte[1], Antoine Zimmermann[2]
Christophe Gravier[1], and Estevam R. Hruschka Jr.[3,4]

[1] Univ Lyon, UJM-Saint-Étienne, CNRS, Laboratoire Hubert Curien
UMR 5516, F-42023 Saint Étienne, France
{jose.gimenez.garcia, maisa.duarte,
christophe.gravier}@univ-st-etienne.fr
[2] Univ Lyon, MINES Saint-Étienne, CNRS, Laboratoire Hubert Curien
UMR 5516, F-42023 Saint-Étienne, France
antoine.zimmermann@emse.fr
[3] Federal University of Sao Carlos - UFSCar, São Carlos, Brazil
[4] Carnegie Mellon University - CMU, Pittsburgh, United States
estevam@cs.cmu.edu

Abstract. NELL is a system that continuously reads the Web to extract knowledge in form of entities and relations between them. It has been running since January 2010 and extracted over 50,000,000 candidate statements. NELL's generated data comprises all the candidate statements together with detailed information about how it was generated. This information includes how each component of the system contributed to the extraction of the statement, as well as when that happened and how confident the system is in the veracity of the statement. However, the data is only available in an ad hoc CSV format that makes it difficult to exploit out of the context of NELL. In order to make it more usable for other communities, we adopt Linked Data principles to publish a more standardized, self-describing dataset with rich provenance metadata.

Keywords: NELL, RDF, Semantic Web, Linked Data, Metadata, Reification

1 Introduction

Never-Ending Language Learning (NELL) [2, 13] is an autonomous computational system that aims at continually and incrementally learning. NELL has been running for about 7 years in Carnegie Mellon University (US). Currently, NELL has collected over 50 million of candidate beliefs, from with about 3.6 million have been promoted as trustworthy statements. NELL learns from the web and uses an ontology previously created to guide the learning. One of the most significant resource contributions of NELL, in addition to the millions of beliefs learned from the Web, is NELL's internal representation (or metadata) for categories, relations and concepts. Such internal representation grows in every iteration, and is used by NELL as a set of different (and constantly updated)

feature vectors to continuously retrain NELL's learning components and build its own way to understand what is read from the Web. Zimmermann et al. [20] published in 2013 a solution to convert NELL's belief into RDF statements. However, NELL's internal representation is not modeled in their work. Thus, the main contribution of this work is to extended the approach to include all the provenance metadata (NELL's internal representation) for each belief. In addition, we publish not only the promoted beliefs, but also the candidates. As far as we know, this dataset contains more metadata about the statements than any other available dataset in the linked data cloud. This in itself can also be interesting for researchers that seek to manage and exploit meta-knowledge. Our intention is to keep this information updated and integrate it on NELL's web page[5].

2 The Never-Ending Language Learning System

NELL [2, 13] was built based on a new Machine Learning (ML) paradigm, the Never-Ending Learning (NEL). NEL paradigm is a semi-supervised learning [1] approach focused on giving the ability to a machine learning system to autonomously use what it has previously learned to continuously become a better learner. NELL is based on a number of coupled components working in parallel. These components read the web and use different approaches to, not only infer new knowledge in the form of beliefs, but also to infer new ways of internally representing the learned beliefs and their properties. Beliefs are divided into candidates and promoted beliefs. In order to be promoted a belief needs to have a confidence score of at least 0.9. Down below we provide a short description of each component:

1. **MBL**is the component responsible for take the promotion decision based on the contributions of the others following components.
2. **CPL** *(Coupled Pattern Learner)* [3] is the component that learns Named Entities (NE) and Textual Patterns (TP) from text in the web pages.
3. **CML** *(Coupled Morphologic Learner)* [3] is responsible for looks at textual regularities and identify.
4. **OE** *(Open Eval)* [17] queries the web and extract small text using predicate instances. OE calculates the score based on the text distance between the instances in a relation.
5. **SEAL** *(Coupled Set Expander for Any Language)* [18] is the component responsible to extract knowledge from HTML patterns. In the past it was called *CSEAL*.
6. **OntologyModifier** is used for any ontology alteration. This component will appear in the Knowledge base when a new seed or and ontology extension is manually introduced.
7. **PRA** *(Path Ranking Algorithm)* [7] is based on Random Walk Inference. PRA analyses the connections between two categories instances which are

[5] http://rtw.ml.cmu.edu

the arguments for a relation. This component replaced the old *Ruler Learner* component.

8. **SpreadsheetEdits** provides modifications in the NELL's Knowledge base using human feedback.
9. **KbManipulation** is used to correct some old bugs from NELL's internal indexing knowledge.
10. **AliasMatcher** find relations between entities and their Wikipedia URL.It is currently not active.
11. **RL** *(Rule Learner)* [10] extracts new knowledge using Horn Clauses based on the ontology. Its implementation was based on FOIL [16] and it is suspended since NELL's Knowledge base was expanded.
12. **LatLong** matches the literal string of Named Entities against a fixed geolocation database.
13. **Semparse** [9] combines syntactic parsing from CCGbank (a conversion of the corpus of trees Penn Treebank [12]) and distant supervision.
14. **LE (Learned Embeddings)** [19] predicts new categories or relations of entities based on Event and Named Entity extraction

3 Converting NELL to RDF

In this section we describe how NELL data and metadata is transformed into RDF. The first subsection presents how NELL's ontology and beliefs are converted, following the work by Zimmermann et al. [20]; the second subsection describes how we convert the provenance metadata associated with each belief. NELL's Knowledge bases used in this paper for the promoted and candidates beliefs are respectively corresponding to the iterations 1055[6] and 1050[7], both are the last versions. The code if publicly available in GitHub[8].

3.1 Converting NELL's beliefs to RDF

NELL's ontology is published as a file with three tab separated values per line, where each line expresses a relationship between categories and other categories, relations, or values used by NELL processes.In order to convert NELL's ontology to RDF each line is transformed into a triple as per Zimmermann et al. [20].

NELL's beliefs are also published in tab-separated format, where each line contains a number of fields to express the belief and the associated metadata, such as iteration of promotion, confidence score, or the activity of the components that inferred the belief. All the fields except 4, 5, 6, and 13 are used to convert the beliefs into RDF statements. For a more detailed description of this step, refer to Zimmermann et al. [20].

[6] http://rtw.ml.cmu.edu/resources/results/08m/NELL.08m.1055.esv.csv.gz
[7] http://rtw.ml.cmu.edu/resources/results/08m/NELL.08m.1050.cesv.csv.gz
[8] https://github.com/WDAqua/nell2rdf

3.2 Converting NELL metadata to RDF

Fields 4, 5, 6, and 13 of each NELL's belief are used to extract the metadata. Each belief is represented by a resource, to which we attach the provenance information. In the promoted beliefs process, field 4 is used to extract the iteration when the belief was promoted, while field 5 gives a confidence score about it. On the other hand, in the candidate beliefs process, fields 4 and 5 contains the iterations when each component generated information about the belief, and the confidence score provided by each of them. Field 6 contains a summary information about the activity of MBL when processing the promoted belief. The complete information from field 6 is presented in the field 13 and because it we are not using field 6. Finally, field 13 is parsed and information about every activity that took part in generating the statement.

The ontology can be seen in Figure 1. We make use of the PROV-O ontology [11] to describe the provenance, and the reification model[9] to represent the beliefs to where we attach the metadata. Each `Belief` can be related with one or more `ComponentIteration` that, in turn, are performed by a `Component`. If the belief is a `PromotedBelief`, it has attached its `iterationOfPromotion` and `probabilityOfBelief`. The `ComponentIteration` is related to information about the process: the `iteration`, `probabilityOfBelief`, `Token`, `source` and `atTime` (the date and time it was processed). The possible values for the `source` object are described in Table 3. The `Token` expresses the concepts that the `Component` is relating. Those concepts can be a pair of entities for a `RelationToken`, and entity and a class for a `GeneralizationToken`, or an entity and values for latitude and longitude for a `GeoToken.`

The classes of the ontology are described in Table 1 and properties of the ontology are described in Table 2.

[9] https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#reification

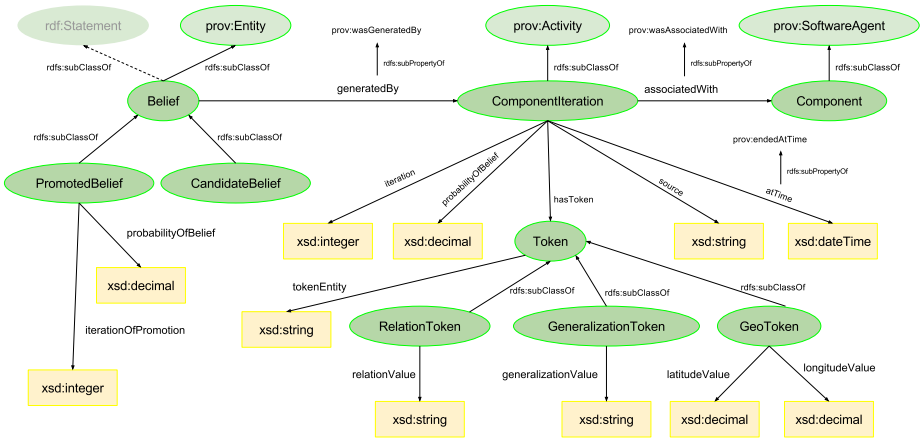

Fig. 1: nellrdf metadata Ontology

Table 1: Description of NELL's metadata ontology classes

Class	rdfs:subClassOf	Description
Belief	prov:Entity	A belief
PromotedBelief	Belief	A promoted belief
CandidateBelief	Belief	A candidate belief
ComponentIteration	prov:Activity	The activity of a component in an iteration
Component	prov:SoftwareAgent	A component
Token	owl:Class	The tuple that was inferred by the activity
RelationToken	Token	The tuple <Entity,Entity> that was inferred for a relation
GeneralizationToken	Token	The tuple <Entity,Category> that was inferred for a generalization
GeoToken	Token	The tuple <Entity,Longitude,Latituded> that was inferred for a geografical belief

Table 2: Description of NELL's metadata properties

Property	rdfs:subPropertyOf	rdfs:domain	rdfs:range
	Description		
generatedBy	prov:wasGeneratedBy	Belief	ComponentIteration
	The Belief was generated by the iteration of the component		
associatedWith	prov:wasAssociatedWith	ComponentIteration	Component
	The iteration was performed by the component		
iterationOfPromotion	owl:DatatypeProperty	PromotedBelief	xsd:integer
	iteration in which the component was promoted		
probabilityOfBelief	owl:DatatypeProperty	PromotedBelief	xsd:decimal
	Confidence score of the Belief		
iteration	owl:DatatypeProperty	ComponentIteration	xsd:integer
	Iteration in which a component performed the activity		
probability	owl:DatatypeProperty	ComponentIteration	xsd:decimal
	Confidence score given by the component		
hasToken	owl:ObjectProperty	ComponentIteration	Token
	The concepts that the component is relating		
source	owl:DatatypeProperty	ComponentIteration	xsd:string
	Data that was used by the component in the activity		
atTime	owl:DatatypeProperty	ComponentIteration	xsd:dateTime
	Date and time when the component iteration was performed		
tokenEntity	owl:DatatypeProperty	Token	xsd:string
	Entity on which the data was inferred		
relationValue	owl:DatatypeProperty	RelationToken	xsd:string
	Entity related the entity appointed by tokenEntity		
generalizationValue	owl:DatatypeProperty	GeneralizationToken	xsd:string
	Class of the entity appointed by tokenEntity		
latitudeValue	owl:DatatypeProperty	GeoToken	xsd:decimal
	Latitude of the entity appointed by tokenEntity		
longitudeValue	owl:DatatypeProperty	GeoToken	xsd:decimal
	Longitude of the entity appointed by tokenEntity		

Table 3: NELL's components and their sources

#	Component Name	Metadata
1	MBL	A string having the entity, relation and value.
2	CPL	A list of textual patterns for category and relation.
3	CMC	A noun phrase list of orthographical features: length and number of words, capitalization, prefixes and suffixes.
4	OE	A list of web text and its URL.
5	SEAL	A list of URL's.
6	OntologyModifier	A string containing the ontology file name. The first file is the ontology used for the system. The second one is the .xls file used to create the ontology one.
7	PRA	A list of the rules.
8	SpreadsheetEdits	A string containing the NE or pair of NE, the action and the feedback file from where it comes. The action will be to promote (marked as +), because when the system has some action is about some,wrong knowledge, this will be removed.
9	KbManipulation	A string containing which part of the ontology (category or relation) was made some correction.
10	AliasMatcher	A date from when the component was executed.
11	RuleInference	A list of rules.
12	LatLong / LatLongTT	A list of latitude and longitude pairs.
13	Semparse	A text.
14	LE	empty

4 The NELL2RDF Dataset

The current version of NELL2RDF updates the promoted beliefs to the last version, adding the provenance triples about them. It also adds the candidate beliefs and their corresponding provenance triples. We provide the dumps for the promoted beliefs [10] and the candidate beliefs[11]. The ontologies for the beliefs[12] and the provenance metadata[13] is common for both dumps. Metadata about the dataset[14] is modeled using VoID and DCAT vocabularies.

5 Discussion and Future Work

In this work we present the conversion of both data and metadata from NELL into RDF. It presents a thesaurus of entities and binary relations between them, as well as a number lexicalizations for each entity. It also includes detailed provenance metadata along with confidence scores.

[10] https://w3id.org/nellrdf/nellrdf.promoted.n3.gz
[11] https://w3id.org/nellrdf/nellrdf.candidates.n3.gz
[12] https://w3id.org/nellrdf/ontology/nellrdf.ontology.n3
[13] https://w3id.org/nellrdf/provenance/ontology/nellrdf.ontology.n3
[14] https://w3id.org/nellrdf/metadata/nellrdf.metadata.n3

Our goals for this dataset are twofold: First, we want improve WDAqua-core0 query answering system, providing it with more relations and lexicalizations, along with confidence scores that can help to give hints about how trustworthy is the answer. Second, given that it contains a big proportion of metadata statements, we want to use it as a testbed to compare how different metadata representations behave. In this respect, the current version makes use of reification to identify each triple, but more datasets with different approaches to represent metadata will be published in the future. Those models include N-Ary properties [15], Quads [4], and Singleton Property [14].

While currently we only publish the dumps of the datasets, we plan to provide SPARQL endpoint and full dereferenceable URLs. In addition, NELL is starting to be explored in languages different than English, such as Portuguese [5, 8] and French [6]. Our intention is to convert those datasets to RDF as they become available to the public.

Acknowledgements: This work is supported by funding from the EU H2020 research and innovation program under the Marie Skłodowska-Curie grant No 642795. We would like to thank Bryan Kisiel from NELL's CMU team for his technical support about NELL's components.

References

[1] Blum, A., Mitchell, T.: Combining Labeled and Unlabeled Data with Co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory (1998)

[2] Carlson, A., Betteridge, J., Hruschka, Jr., E.R., Mitchell, T.M.: Coupling Semi-supervised Learning of Categories and Relations. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing (2009)

[3] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr, E.R.H., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010) (2010)

[4] Carothers, G.: RDF 1.1 N-Quads: A line-based syntax for RDF datasets. W3C Recommendation (2014), https://www.w3.org/TR/n-quads

[5] Duarte, M.C., Hruschka, E.R.: How to Read The Web In Portuguese Using the Never-Ending Language Learner's Principles. In: Proceedings of the 14th International Conference on Intelligent Systems Design and Applications (2014)

[6] Duarte, M.C., Maret, P.: Vers une instance française de NELL : Chaîne TLN multilingue et modélisation d'ontologie. Revue des Nouvelles Technologies de l'Information (2017)

[7] Gardner, M., Talukdar, P.P., Krishnamurthy, J., Mitchell, T.M.: Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases. In: Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of the ACL (2014)
[8] Hruschka, E.R., Duarte, M.C., Nicoletti, M.C.: Coupling as Strategy for Reducing Concept-Drift in Never-ending Learning Environments. Fundamenta Informaticae (1) (2013)
[9] Krishnamurthy, J., Mitchell, T.M.: Joint Syntactic and Semantic Parsing with Combinatory Categorial Grammar. In: ACL (2014)
[10] Lao, N., Mitchell, T., Cohen, W.W.: Random Walk Inference and Learning in a Large Scale Knowledge Base. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2011)
[11] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The prov ontology. W3C Recommendation (2013), https://www.w3.org/TR/prov-o/
[12] Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The Penn Treebank: Annotating Predicate Argument Structure. In: Proceedings of the Workshop on Human Language Technology (1994)
[13] Mitchell, T.M., Cohen, W.W., Hruschka, Jr., E.R., Talukdar, P.P., Betteridge, J., Carlson, A., Mishra, B.D., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E.A., Ritter, A., Samadi, M., Settles, B., Wang, R.C., Wijaya, D.T., Gupta, A., Chen, X., Saparov, A., Greaves, M., Welling, J.: Never-Ending Learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. (2015)
[14] Nguyen, V., Bodenreider, O., Sheth, A.P.: Don't like RDF reification?: Making statements about statements using singleton property. In: 23rd International Conference on World Wide Web (2014)
[15] Noy, N., Rector, A., Hayes, P., Welty, C.: Defining n-ary relations on the semantic web. W3C Working Group Note (4) (2006), https://www.w3.org/TR/swbp-n-aryRelations/
[16] Quinlan, J.R., Cameron-Jones, R.M.: FOIL: A Midterm Report. In: Proceedings of the European Conference on Machine Learning (1993)
[17] Samadi, M., Veloso, M.M., Blum, M.: OpenEval: Web Information Query Evaluation. In: AAAI (2013)
[18] Wang, R.C., Cohen, W.W.: Language-Independent Set Expansion of Named Entities Using the Web. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (2007)
[19] Yang, B., Mitchell, T.: Joint Extraction of Events and Entities within a Document Context. In: NAACL (2016)
[20] Zimmermann, A., Gravier, C., Subercaze, J., Cruzille, Q.: Nell2RDF: Read the Web, and Turn it into RDF. In: KNOW@ LOD (2013)