Curating an Open Information Extraction Knowledge Base Using Games with a Purpose

Kevin Forand
RALI/DIRO
University of Montreal
CP 6128. Succ. Centre-Ville
H3C3J7 Montréal, Québec, Canada
skulg@hotmail.com

Philippe Langlais
RALI/DIRO
University of Montreal
CP 6128. Succ. Centre-Ville
H3C3J7 Montréal, Québec, Canada
felipe@iro.umontreal.ca

Abstract

We are interested in measuring how games with a purpose can be used as a crowdsourcing solution for transforming a (huge) set of triples extracted from Wikipedia into a useful knowledge base. We describe the natural language processing pipeline used for generating questions that we turned into games. We present three games we implemented.

1 Introduction

Open information extraction (OIE) (Banko et al., 2007) is a powerful tool that can help with the task of building a large knowledge base. By analyzing large quantities of texts (for instance the Web), OIE tools can help collect a huge collection of triples such as *(Chilly_Gonzales,is_a,pianist)*. This extraction process is however highly noisy, and many triples acquired that way are either uninformative or invalid statements. Therefore, a validation phase is recommended; but manually curating a collection of triples is an exceptionally intimidating task. In this work, we investigate the use of games with a purpose (Von Ahn and Dabbish, 2008) as a crowdsourcing method to reduce the time involved in sanitizing the knowledge base.

2 Related Work

Open information extraction techniques have been used to gather large amounts of triples from text. Several OIE extractors are nowadays available, including those developed at University of Washington,[1] `Reverb` (Fader et al., 2011) being one of the earliest one. Organizing a collection of triples into a useful database has been the focus of many studies, including the work conducted within the Never Ending Language Learning project (Carlson et al., 2010) or the impressive deployment done at Google for collecting their *Knowledge Vault* database (Dong et al., 2014).

As demonstrated by the `Verbosity` (Von Ahn et al., 2006) and the `JeuxDeMots` (Lafourcade and Joubert, 2013) projects, games with a purpose (`GwaPs`) can be quite successful as a crowdsourcing tool. Several `GwaPs` have been proposed to help acquire or curate knowledge. For instance Kumaran et al. (2014) implemented a pictionary inspired game for acquiring textual paraphrases, while in (Vannella et al., 2014) the authors investigate the use of dynamic gameplays for curating lexical knowledge.

3 Knowledge Base

The point of departure in this work is a set of over 30 millions of triples *(arg1,relation,arg2)* collected from the French part of Wikipedia, thanks to an adaptation of `Reverb` into French. This work is described in (Gotti and Langlais, 2016).

3.1 Noise filtration

We applied a number of filters to more or less aggressively remove spurious triples. We removed those that had an hapax argument (arg1 or arg2) in the collection, those containing special symbols such as mathematical ones. More importantly, we removed as well triples with pronouns, *e.g. (She,is_mother_of,Elisabeth)*, which are typically useless in the absence of a good coreference resolution chain. Overall, we eliminated more than half of the triples that way.

3.2 Category assignation

We defined a set of categories by searching for the pattern *(arg1,is,a arg2)* in our triple store, and kept the 1 000 most frequent arg2 as categories. This

[1] `https://www.cs.washington.edu/node/3540`

simple process avoids to rely on a manually defined set of categories, and can potentially be applied on any other triple store. For instance, we have categories such as *Writer*, *Musician*, *Place* or *Problem*.[2]

3.3 Relational Profiles

We computed for each category its *relational profile*, that is, the distribution of relations entities of the category typically interact with. For instance *is*, *has*, *produces*, *records*, *becomes*, *participates* are the most likely relations that characterize the category *Singer*. We first computed a relational profile of each arg1 in the triple store. Then, we aggregated the profiles of the terms we could associate to a category by the pattern previously described. Since the relation *is* is present in the relational profile of all the categories, we decided to remove this relation (and renormalized).

3.4 Similarities

We calculated the similarities between the 100 000 most frequent arguments (arg1) and the 1 000 categories with the cosine similarity measure applied to the relational profiles previously computed. Those similarity scores help us to extend the set of likely categories of a given term, upon the category attributed by our pattern-based approach. For instance, the term *Mozart* was not initially associated to any category by our patterns, but thanks to similarity scores, we hypothesize *Composer*, *Poet*, *Song-writer*, *Musician*, *Writer* as likely ones.

4 Gamification

We developed a series of games to help curate our triple store. We identified three tasks that we transformed into games we implemented using the `Unity` platform[3]; a plateform dedicated to the development of 2D and 3D animated games. The games are currently being tested internally.

4.1 Classification

The task of classification helps us validate the category assigned to a term (arg1) either by our pattern-based approach or with the similarities computed. A typical quiz is illustrated in Figure 1a. We designed two gameplay for this. The first one involves a pinball table where terms are the balls that a user has to send to specific zones that are associated to categories. A second version offers a simplified gameplay where the user simply has to drag the terms towards the correct categories plotted on the screen. While the former game involves more dexterity, and can therefore lead to noisy output, we expect it to be more attractive.

a) Associate terms to categories:
terms blue, painting, money, religion
categories Picture, Art, Color, *others*

b) Search for intruders of *Singer*:
Félix Leclerc, Montréal, Ben Zimet
Duo, Francis Cabrel

c) Complete the missing argument:
Singer, sings, ?
Anders_Fridén, records, ?

Figure 1: Examples of quizzes automatically generated for our games.

4.2 Filtering

In this task, the user is asked to assert certain facts. We adapted it into a gameplay where a user has to find intruders among a set of terms and a given category. A typical quiz is illustrated in Figure 1b.

4.3 Facts Discovery

The goal of this task is for the user to complete some statements (triples) in which part of the information is hidden (currently arg2). This allows to extend the KB with knowledge that was unknown at extraction time. Our gamification enforces a user to drag letters into forming a term. An example of a typical quiz is illustrated in Figure 1c.

5 Discussion

We intend to announce those games during March 2017, and are eager to measure how the output collected can help us curate our triple store. The games and more information will be available at *http://rali.iro.umontreal.ca/rali/en/oie-kb-using-gwap*.

Acknowledgments

This work is part of the TRiBE project founded by NSERC.

[2]Examples are translated into English.
[3]`https://unity3d.com/`

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610. ACM.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545.

Fabrizio Gotti and Philippe Langlais. 2016. Harnessing open information extraction for entity classification in a french corpus. In *Canadian AI 2016*. Springer International Publishing Switzerland, Springer International Publishing Switzerland.

A Kumaran, Melissa Densmore, and Shaishav Kumar. 2014. Online gaming for crowd-sourcing phrase-equivalents. In *25th International Conference on Computational Linguistics: Technical Papers*, pages 1238–1247, August.

Mathieu Lafourcade and Alain Joubert. 2013. Bénéfices et limites de l'acquisition lexicale dans l'expérience jeuxdemots. *Ressources Lexicales: Contenu, construction, utilisation, évaluation, LingvisticæInvestigationes, Supplementa*, 30:187–216.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 1294–1304.

Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78. ACM.