Empirical Analysis of Foundational Distinctions in Linked Open Data

Luigi Asprino[1,2]**, Valerio Basile**[3]**, Paolo Ciancarini**[2] **and Valentina Presutti**[1]

[1] STLab, ISTC-CNR, Rome, Italy

[2] University of Bologna, Bologna, Italy

[3] University of Turin, Turin, Italy

luigi.asprino@unibo.it, basile@di.unito.it, paolo.ciancarini@unibo.it, valentina.presutti@cnr.it

Abstract

The Web and its Semantic extension (i.e. Linked Open Data) contain open global-scale knowledge and make it available to potentially intelligent machines that want to benefit from it. Nevertheless, most of Linked Open Data lack ontological distinctions and have sparse axiomatisation. For example, distinctions such as whether an entity is *inherently* a class or an individual, or whether it is a physical object or not, are hardly expressed in the data, although they have been largely studied and formalised by foundational ontologies (e.g. DOLCE, SUMO). These distinctions belong to common sense too, which is relevant for many artificial intelligence tasks such as natural language understanding, scene recognition, and the like. There is a gap between foundational ontologies, that often formalise or are inspired by pre-existing philosophical theories and are developed with a top-down approach, and Linked Open Data that mostly derive from existing databases or crowd-based effort (e.g. DBpedia, Wikidata). We investigate whether machines can learn foundational distinctions over Linked Open Data entities, and if they match common sense. We want to answer questions such as "does the DBpedia entity for *dog* refer to a class or to an instance?". We report on a set of experiments based on machine learning and crowdsourcing that show promising results.

1 Common Sense and Linked Open Data

Common Sense Knowledge is knowledge about the world, shared by all people. Common sense reasoning is also at the core of many Artificial Intelligence (AI) tasks such as natural language understanding, object and action recognition, and behavior arbitration [Davis and Marcus, 2015], but it is difficult to teach to those systems. Its importance was assessed back in 1989 by [Hayes, 1989] who argued that AI needs a "formalization of a sizable portion of commonsense knowledge about the everyday physical world" (cit.), which, he says, must have three main characteristics: uniformity, density, and breadth. The Semantic Web effort has partly addressed this requirement with Linked Open Data (LOD):

∼150 billion linked facts[1], formally and uniformly represented in RDF and OWL, and openly available on the Web. Nevertheless, LOD still fails in addressing density (high ratio of facts about concepts) and breadth (large coverage of physical phenomena). In fact, it is very rich for domains such as geography, linguistics, life sciences and scholarly publications, as well as for cross-domain knowledge, but it mostly encodes this knowledge from an encyclopaedic perspective. The ultimate goal of our research is to enrich LOD with common sense knowledge, going beyond the encyclopaedic view. We claim that an important gap to be filled towards this goal is: assessing foundational distinctions over LOD entities, that is to distinguish and formally assert whether a LOD entity inherently refers to e.g. a class or an individual, a physical object or not, a location, a social object, etc., from a common sense perspective.

1.1 Foundational Distinctions

High level categorial distinctions (e.g. class vs. instance) are a fundamental human cognitive ability: "There is nothing more basic than categorization to our thought, perception, action, and speech." [Lakoff, 1987]. This is also why "the organisation of objects into categories is a vital part of knowledge representation" [Russell and Norvig, 2009]. Foundational distinctions have been theorised and modelled in foundational ontologies such as DOLCE [Masolo *et al.*, 2003] and SUMO [Pease and Niles, 2002] with a top-down approach, but populating and empirically validating them has been rarely addressed. In this study, we perform a set of experiments to assess *whether machines can learn to perform foundational distinctions, and if they match common sense*. The former issue is investigated by applying machine learning techniques, the latter is assessed by crowdsourcing the annotation of the evaluation datasets. We use DOLCE+DnS UltraLite (DUL)[2] as reference foundational ontology to select the targets of our experiments. We start by focusing on two very basic but diverse distinctions, which need to be addressed before approaching all the others: whether a LOD

---

[1] http://stats.lod2.eu/, accessed on April 25th 2018

[2] DOLCE+DnS UltraLite (DUL) http://www.ontologydesignpatterns.org/ont/dul/DUL.owl is derived from DOLCE. DUL inherits most of DOLCE's distinctions by providing a more intuitive terminology and simplified axiomatisation. It has been widely adopted by the Semantic Web community.

entity e.g. `dbr:Rome`[3], (i) inherently refers to a class or an instance, and whether it (ii) refers to a physical object or not. The first distinction (class vs. instance) is fundamental in formal ontology, as evidenced by upper-level ontologies (e.g. SUMO and DOLCE), and showed its practical importance in modelling and meta-modelling approaches in computer science, e.g. the class/classifier distinction in Meta Object Facility[4]. It is also at the basis of LOD knowledge representation formalisms (RDF and OWL) for supporting taxonomic reasoning (e.g. inheritance). Automatically learning whether a LOD entity is a class or an instance – from a common sense perspective – impacts on the behaviour of practical applications relying on LOD as common sense background knowledge. Examples include: question answering, knowledge extraction, and more broadly human-machine interaction. In fact, many LOD datasets that are commonly used for supporting these tasks (especially general purpose datasets e.g. DBpedia, Wikidata, BabelNet) only partially, and often incorrectly, assert whether their entities are classes or instances, and this has been proved to be a source of many inconsistencies and error patterns [Paulheim and Gangemi, 2015].

The second distinction (physical object or not) is essential to represent the physical world. In fact, only physical objects can move in space or be the subject of axioms expressing their expected (naive) physical behaviour (e.g. gravity). We refer to the definition of *Physical Object* provided by DUL: *"Any Object that has a proper space region. The prototypical physical object has also an associated mass, but the nature of its mass can greatly vary based on the epistemological status of the object (scientifically measured, subjectively possible, imaginary)"*.

1.2 Contribution

In the reminder of this paper we describe an automated way of making these distinctions emerge empirically. We present and discuss a set of experiments, conducted on a sample of LOD, involving manual inspection, ontology alignment, machine learning, and crowdsourcing. The obtained results are promising, and motivate us to extend them to a much larger scale, in line with a recent inspiring talk "The Empirical Turn in Knowledge Representation"[5] by van Harmelen, who suggests that LOD is a unique opportunity to "observe how knowledge representations behave at very large scale".

In summary, the contribution of this research includes:

- a novel method that leverages supervised machine learning and crowdsourcing to automatically assess foundational distinctions over LOD entities (cf. Section 3), according to common sense;

- four reusable datasets, based on a sample of DBpedia, separately annotated by experts and by the crowd with class/instance and physical object classification, for each entity (cf. Section 4). The crowdsourced task designs are on their turn reusable;

- a set of reproducible experiments targeting two foundational distinctions: *class vs. instance* and *physical object vs. not a physical object*, showing that machines can learn them by using a same set of features, and that they match common sense (cf. Section 5).

2 Related Work

Only a few studies focus on typing entities based on **foundational distinctions**. To the best of our knowledge, our research is the first to test the hypothesis that machines can learn foundational distinctions that match common sense, by using web resources. The closest work to ours in approach and scale is [Gangemi *et al.*, 2012], which produced a dataset of DBpedia entities annotated with DUL classes, using ontology learning. We reuse this dataset and compare our results against it. The work by [Miller and Hristea, 2006] addresses the problem of distinguishing classes from instances in WordNet [Fellbaum, 1998] synsets, through purely manual annotation. This approach is inappropriate to test our research questions due to its lack of scalability. Over the years, a number of **common sense knowledge bases** have been proposed for supporting diverse tasks spanning from automated reasoning to natural language processing. We decided to use the English DBpedia[6] [Bizer *et al.*, 2009] in our experiments. Besides being very popular, it is the *de facto* main hub of LOD, with its 4.58 million entities[7]. There are other resources that either target common sense, or encode potentially relevant knowledge for common sense reasoning. However they lack explicit assertions of foundational distinctions, and show very sparse coverage distribution. Two of them are relevant in this context, and we plan to investigate how to leverage them in future experiments that we plan to perform at a much larger scale: ConceptNet[8] [Liu and Singh, 2004] is a large-scale semantic network that integrates other existing resources, mostly derived from crowdsourced resources, expert-created resources, and games with a purpose. We found it unsuitable for reuse at this stage of our study: only a very small portion of it is linked to LOD, and its representation is bound to linguistic expressions instead of formalised concepts. OpenCyC[9] [Lenat, 1995] includes an ontology and a knowledge base of common sense knowledge organised in modular theories, released as part of the long-standing CyC project. OpenCyC has been mainly developed with a top-down approach by ontology engineering experts. It is a potentially valuable resource to compare with, but it uses a proprietary representation language, which makes it hard to work with, currently. There is a plan to make it available to the scientific community as linked data, although apparently inaccessible at the time of our experiment.

---

[3] `dbr:` stands for http://dbpedia.org/resource/

[4] https://www.omg.org/spec/MOF/

[5] https://goo.gl/BDSGY1

[6] http://dbpedia.org

[7] http://wiki.dbpedia.org/about/facts-figures, accessed on April 25th 2018

[8] http://conceptnet5.media.mit.edu/

[9] http://www.opencyc.org/

3 Automatic Classification of Foundational Distinctions

We want to answer the following research questions: *(RQ1) Do foundational distinctions match common sense? (RQ2) Does the (Semantic) Web provide an empirical basis for supporting foundational distinctions over LOD entities, according to common sense? and (RQ3) what ensemble of features, resources, and methods works best to make machines learn foundational distinctions over LOD entities?*

Our objective is to test all the distinctions formalised in DUL. Nevertheless, for each of them we need to create a set of reference datasets (cf. Section 4) in order to train and evaluate the proposed method, which requires a significant amount of work. For this reason, as anticipated in Section 1, we start focusing on two distinctions: between class and instance, and between what is a physical object and what is not. These are two of the most basic, but very diverse distinctions in knowledge representation and formal ontology. The former applies at a very high level, and is usually modelled by means of logical language primitives (e.g. `rdf:type`, `rdfs:subClassOf`). The latter concerns the identification of those entities that constitute the physical world, hence highly relevant and primitive as far as common sense about physics is concerned. We argue that by investigating these two distinctions, given their diverse character, we can assess the feasibility of a larger study based on the proposed method, and get an indication of its generalisability. We use DBpedia (release 2016-10) in our study as most LOD datasets link to it. We approach this problem as a classification task, using two classification approaches: *alignment-based* (cf. Section 3.1) and *machine learning-based* (cf. Section 3.2). Since no established procedure exists, we tested different families of methods in an exploratory way. This led us to reuse – or compare to – existing work, which provides us with a baseline, which includes Tìpalo [Gangemi *et al.*, 2012] as well as other relevant alignments between DBpedia and lexical resources (cf. Section 3.1).

3.1 Alignment-based Classification

Alignment-based methods exploit the linking structure of LOD, in particular the alignments between DBpedia, foundational ontologies, and lexical linked data, i.e. LOD datasets that encode lexical/linguistic knowledge. The advantage of these methods is their inherent unsupervised nature. Their main disadvantages are the need of studying the data models for designing suitable queries, and the potential limited coverage and errors that may accompany the alignments. We have developed **SENECA** (Selecting Entities Exploiting Linguistic Alignments), which relies on existing alignments in LOD, to make an automatic assessment of the foundational distinctions asserted over DBpedia entities. A graphical description of SENECA is depicted in Figure 1.

Class vs. Instance. As far as this distinction is concerned, SENECA works based on the hypothesis that common nouns are mainly classes and they are expected to be found in dictionaries, while it is less the case for proper nouns, that usually denote instances. This hypothesis was suggested by [Miller and Hristea, 2006], who manually annotated in-
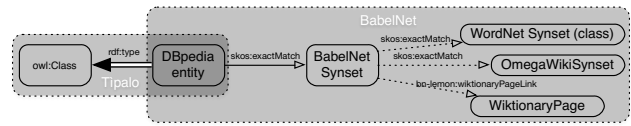
(a) The alignment paths followed by SENECA for selecting candidate classes among DBpedia entities. It identifies as classes all DBpedia entities aligned via BabelNet to a WordNet synset, an OmegaWiki synset or a Wiktionary page, and all DBpedia entities typed as `owl:Class` in Tìpalo.
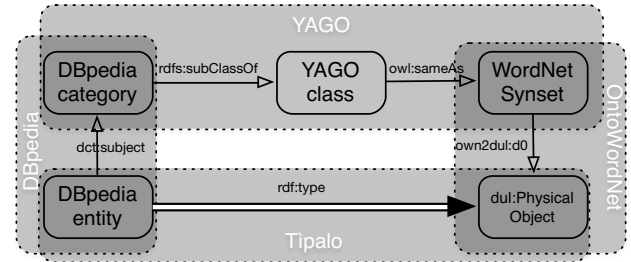
(b) The alignment paths used by SENECA for identifying candidate Physical Objects among DBpedia entities. It navigates the YAGO taxonomy that via OntoWordNet links DBpedia entities to `dul:PhysicalObject` or Tìpalo that links DBpedia entities to `dul:PhysicalObject`.

Figure 1: SENECA approach for assessing whether a DBpedia entity is a class or an instance (Figure 1a) and whether it is a physical object or not (Figure 1b).

stances in WordNet, information that SENECA reuses when available. A good quality alignment between the main LOD lexical resources and DBpedia is provided by BabelNet [Navigli and Ponzetto, 2012][10]. SENECA exploits these alignments and selects all the DBpedia entities that are linked to an entity in WordNet[11], Wiktionary[12] or OmegaWiki[13]. With this approach, 63,620 candidate classes have been identified, as opposed to WordNet annotations that only provide 38,701 classes. In order to further increase the potential coverage, SENECA leverages the typing axioms of Tipalo [Gangemi *et al.*, 2012], broadening it to 431,254 total candidate classes. All the other DBpedia entities are assumed to be candidate instances. SENECA criteria for selecting candidate classes among DBpedia entities are depicted in Figure 1a.

Physical Object. Almost 600,000 DBpedia entities are only typed as `owl:Thing` or not typed at all. However, each DBpedia entity belongs to at least one Wikipedia category. Wikipedia categories have been formalised as a taxonomy of classes (i.e. by means of `rdfs:subClassOf`) and aligned to WordNet synsets in YAGO [Suchanek *et al.*, 2007][14]. WordNet synsets are in turn formalised as an OWL ontology

---

[10]We use BabelNet 3.6, which is aligned to WordNet 3.1

[11]http://wordnet-rdf.princeton.edu/, we use WordNet 3.0 and its alignments to WordNet 3.1, to ensure interoperability with the other resources

[12]https://www.wiktionary.org/

[13]http://www.omegawiki.org/

[14]We use YAGO 3, aligned to WordNet 3.1

in OntoWordNet [Gangemi *et al.*, 2003][15]. OntoWordNet is based on DUL, hence it is possible to navigate the taxonomy up to the DUL class for Physical Object. SENECA looks up the Wikipedia category of a DBpedia entity and follows these alignments. Additionally, it uses Tìpalo, which includes type axioms of DBpedia entities based on DUL classes. SENECA uses these paths of alignments and taxonomical relations, as well as the automated inferences that enable to assess whether a DBpedia entity is a Physical Object or not. With this approach, graphically summarised in Figure 1b, 67,005 entities were selected as candidate physical objects.

3.2 Machine Learning-based Classification

Within machine learning, *classification* is the problem of predicting which category an entity belongs to, given a set of examples, i.e. a training set. The training set is processed by an algorithm in order to learn a predictive model based on the observation of a number of *features*, which can be categorical, ordinal, integer-valued or real-valued. We have designed our target distinctions in the form of two binary classifications. We have experimented with eight classification algorithms: J48, Random Forest, REPTree, Naive Bayes, Multinomial Naive Bayes, Support Vector Machines, Logistic Regression, and K-nearest neighbours classifier. We have used WEKA[16] for their implementation.

Features. The classifiers were trained using the following four features.

Abstract. Considering that DBpedia entities are all associated with an abstract providing a definitional text, our assumption is that these texts encode useful distinctive patterns. Hence, we retrieve DBpedia entity abstracts, and represent them as 0-1 vectors (bags of words). We built a dictionary containing the 1000 most frequent tokens found in all the abstracts of the dataset. The dictionary is case-sensitive, since the tokens are not normalised. The resulting vector has a value 1 for each token mentioned in the abstract, 0 for the others. By inspecting a good amount of abstracts, we noticed that very frequent words, such as conjunctions and determiners, are used in a way that can be informative for this type of classifications. For example, most of class definitions begin with "A" ("A knife is a tool..."). For this reason, we did not remove stop-words.

URI. We notice that the ID part of URIs is often as informative as a label, and often follows conventions that may be discriminating especially for the class vs. instance classification. In DBpedia, the ID of a URI reflects an entity name (it is common practice in order to make the URI more human-readable), and it always starts with an upper case letter. If the entity's name is a compound term and the entity denotes an instance, each of its components starts with a capital letter. We have also noticed that DBpedia entity names are always mentioned at the beginning of their abstract and, for most of the instance entities, they have the same capitalisation pattern as the URI ID. Moreover, instances tend to have more terms in their ID than classes. These observations were captured by three numerical features: (i) number of terms in the ID start-

ing with a capital letter, (ii) number of terms in the ID that are also found in the abstract, and (iii) number of terms in the ID.

Incoming and Outgoing Properties. As part of our exploratory approach, we want to test the ability of LOD to show relevant patterns leading to foundational distinctions. Given that triples are the core tool of LOD, we model a feature based on ongoing and outgoing properties of a DBpedia entity. An outgoing property of a DBpedia entity is a property of a triple having the entity as subject. On the contrary, an incoming property is a property of a triple having the entity as object. For example, considering the triple `dbr:Rome :locatedIn dbr:Italy`, the property `:locatedIn` is an outgoing property for `dbr:Rome` and an incoming property for `dbr:Italy`. For each DBpedia entity, we count its incoming and outgoing properties, per type. For example, properties such as `dbo:birthPlace` or `dbo:birthDate` are common outgoing properties of an individual person, hence their presence suggests that the entity is an individual.

Outcome of SENECA. Following an exploratory approach, we decided to use the output of SENECA as a binomial feature (taking value "yes" or "no") for the classifiers (excluding Multinomial Naive Bayes classifier).

4 Reference Datasets for Classification Experiments

In order to perform our experiments and evaluate the results of our approach (cf. Section 3), we have created two datasets for each of the foundational distinctions under study: one annotated by experts, and another one annotated by the crowd. In this way we can get an indication whether foundational distinctions match common sense (cf. *RQ1*). The resulting datasets include a sample of annotated DBpedia entities and are available online [17].

Selecting DBpedia Entities. The first step to build the datasets is to select a sample of DBpedia entities to be submitted for annotations. It is not straightforward to select a balanced number of classes and instances from DBpedia. A random selection would cause a strong unbalance towards instances because DBpedia contains a larger number of named individuals – e.g. Rome or Barack Obama – than concepts. A possible solution could be to manually select a sufficient equal number of DBpedia instances and classes, however this may inject a bias in the datasets. We have opted for a compromise solution by following the intuition that if entities are represented as vectors, their neighbour vectors would include classes and instances with a more balanced ratio than the random choice. As for minimising the bias, we only manually select an arbitrarily small (i.e. 20) number of seeds (equally distributed). We leverage NASARI [Camacho-Collados *et al.*, 2016], a resource of vector representations for BabelNet synsets and Wikipedia entities. For each vector corresponding to a seed entity, we retrieve its 100 nearest neighbours[18]. After cleaning off duplicated entities (e.g. Wikipedia redirects), entities without abstracts, disambiguation pages, etc.,

---

[15]OntoWordNet is aligned to WordNet 3.0

[16]https://www.cs.waikato.ac.nz/ml/weka/

[17]https://github.com/fdistinctions/ijcai18

[18]We observed that by picking 100 nearest neighbour entities the cosine similarity is always above a threshold of 0.6

we still assessed (through expert annotations) an unbalance towards instances. In the light of a broader usage of the same dataset to also annotate the distinction between physical objects and not physical objects, we enriched the sample with class entities representing physical locations (a good source of physical objects). In order to select only physical location classes from DBpedia, we used the corresponding DBpedia category `dbc:Places` and the SUN database[19], a computer vision-oriented dataset containing a collection of annotated images, covering a large variety of environmental scenes, places, and the objects within. We retrieved DBpedia entities whose labels match SUN locations or that belong to the category `dbc:Places`, and added a number of them to the sample that would suffice to improve the balance. As a result, a total number of 4502 entities were collected in the newly created dataset.

CI$_E$ Dataset: Class vs. Instance Annotation Performed by Experts. Two authors of the paper have manually and independently annotated all entities by indicating whether they were instances or classes, using the associated DBpedia abstract as reference description. Their judgements showed an agreement of 93,6%: they only disagreed on 286 entities. From a joint second inspection, they agreed on additional 281 entities that were initially misclassified by one of the two. Examples of misclassified entities are: `dbr:Select_Comfort` (a U.S. manufacturer) that was erroneously annotated as class; `dbr:Catawba_Valley_Pottery` (a kind of pottery) annotated as instance instead of class. Among the remaining entities, five are *polysemous* cases, where the entity and its description point to both types of referents, e.g. `dbr:Slide_Away_Bed` is a trademark commonly used also to refer to a type of beds. The authors decided to annotate these entities as classes. As a result, the CI$_E$ annotated dataset contains 1983 classes and 2519 instances, which is reasonable balanced (44% classes, 56% instances).

CI$_C$ Dataset: Class vs. Instance Annotation Performed by Crowd. The same dataset was then crowdsourced: each worker was asked to indicate whether an entity is a class or an instance based on its name, abstract, and its corresponding Wikipedia article. We want to assess the agreement between the experts and the crowd, which indicates whether foundational distinctions match common sense or not (cf. *RQ1*). The task was executed on Figure Eight [20] by English speakers with high trustworthiness. The quality of the contributors has been assessed with 51 test questions with a tolerance of only 20% of errors. We collected 22,510 judgments from 117 contributors: each entity was annotated by at least 5 different workers. For each entity e, we computed the level of agreement on each class c, weighted by the trustworthiness scores

$$agreement(e, c) = \frac{SumOfTrust(e, c)}{SumOfTrustOfWorkers(e)} \quad (1)$$

where $SumOfTrust(e, c)$ is the sum of the trustworthiness scores of the workers that annotated entity e with class c; and $SumOfTrustOfWorkers(e)$ is the sum of the trust-

[19]https://groups.csail.mit.edu/vision/SUN/

[20]https://www.figure-eight.com/

Agreement	# Class	# Instance	Total
≥ 0.5	1934	2568	4502
≥ 0.6	1884	2495	4379
≥ 0.8	1631	2330	3961

Table 1: CI$_C$ dataset crowd-based annotated dataset of classes and instances. The table provides an insight of the dataset per level of agreement. Agreement values computed according to Formula 1.

worthiness scores of all the workers that annotated the entity e. Table 1 reports the results of the task indicating the distribution of classes and instances per level of agreement. The average agreement of the crowd is 95.76% . We compared crowd's annotations (with agreement greater than 0.5) against experts' ones. The judgements of the crowd workers diverge from the experts' only on 193 entities, i.e. agreement is 95,7%, suggesting that *the instance vs. class foundational distinction matches common sense* (cf. RQ1 applied to this distinction). Some of those entities (61) also caused a disagreement between experts, hence denoting ambiguous cases. Examples include polysemic entities such as `dbr:Zeke_the_Wonder_Dog` or music genres (e.g. `dbr:Ragga`).

PO$_E$ Dataset: Physical Object Annotation Performed by Experts. Two authors of the paper further annotated (independently) the dataset by indicating for each entity whether it referred to a physical object (PO) or not (NPO), using its DBpedia abstract as reference description. They only disagreed on 272 entities, showing an agreement of 93,9%. By means of a joint second inspection, they agreed that the disagreement was caused by errors in the classification, some of which were borderline cases e.g.: *communities* (e.g. `dbr:Desert_Lake,_California`), wrongly interpreted as society instead of neighbourhood, *trademarked materials* (e.g. `dbr:Waxtite`) and entities with complex description (e.g. `dbr:Cabaña_pasiega`). The resulting PO$_E$ annotated dataset contains 3055 POs and 1447 NPOs.

PO$_C$ Dataset: Physical Object Annotation Performed by Crowd. We also crowdsourced the annotations of physical objects vs. not a physical object: the workers were asked to perform this task by using the entity's name, its abstract, and Wikipedia page as reference descriptions. The quality of the workers has been assessed with 49 test questions, used to exclude contributors that scored an accuracy lower than 80% We collected 25,776 judgments from 173 workers. Each entity has been annotated by at least 5 different English speakers. Table 2 summarises the level of agreement associated with the distribution of PO vs. NPO annotations. The average agreement of the crowd's annotations is 85.48% . The agreement between the crowd and the experts is 85,69%, suggesting that *the PO vs NPO foundational distinction also matches common sense* (cf. RQ1 applied to this distinction).

5 Experiments Results and Discussion

The results of the performed experiments are expressed in terms of precision, recall and F1 measure, computed for each classification and for each target class (class vs. instance and physical object vs. ¬ physical object). The average F1 score

Agreement	# Physical Object	# ¬ Physical Object	Total
≥ 0.5	3601	901	4502
≥ 0.6	3448	641	4089
≥ 0.8	2989	335	3324

Table 2: PO_C dataset: crowd-based annotated dataset of physical objects. The table provides an insight of the dataset per level of agreement. Agreement values computed according to Formula 1.

Dataset	P^C	R^C	F_1^C	P^I	R^I	F_1^I	$\overline{F_1}$
CI_E	.919	.693	.796	.753	.939	.836	.813
CI_C	**.935**	**.731**	**.818**	**.778**	**.945**	**.853**	**.836**

Dataset	P^{PO}	R^{PO}	F_1^{PO}	P^{NPO}	R^{NPO}	F_1^{NPO}	$\overline{F_1}$
PO_E	.877	.247	.385	.561	.965	.713	.548
PO_C	**.954**	**.247**	**.393**	**.567**	**.988**	**.721**	**.557**

Table 3: Results of SENECA on the Class vs. Instance and Physical Object classifications compared against the reference datasets described in Section 4. P^*, R^* and F_1^* indicate precision, recall and F1 measure on Class (C), Instance (I), Physical Object (PO) and complement of Physical Object (NPO). $\overline{F_1}$ is the average of the F1 measures.

is also provided. We compare the results of the methods, described in Section 3, against the reference datasets CI_E, CI_C, PO_E and PO_C, described in Section 4. As for CI_C and PO_C, we only include the annotations having agreement $\geq 80\%$.

5.1 Alignment-based Methods: SENECA

Class vs. Instance. Table 3 shows SENECA's performance on the class vs. instance classification, by comparing its results with CI_E and CI_C. SENECA shows very good performance with best avg F1 $= .836$, when compared with CI_C. Considering that SENECA is unsupervised, and is based on existing alignments in LOD, this result suggests that LOD may better reflect common sense than the expert's perspective, an interesting hint for further investigation on this specific matter.

Physical Object. Table 3 shows the performance of SENECA on the Physical Object classification task computed by comparing its results with PO_E and PO_C (cf. Section 4). We observe a significant drop in the best average F1 score (.557) as compared to the class vs. instance classification task (.836). On one hand, this may suggest that the task is harder. On the other hand, the alignment paths followed in the two cases are different, since for classifying Physical Objects more alignment steps are required. In the first case (class vs. instance), BabelNet directly provides the final alignment step (cf. Figure 1a), while in the second case (PO vs. NPO), three more alignment steps are required: DBpedia Category $\rightarrow$ YAGO $\rightarrow$ WordNet (cf. Figure 1b). It is reasonable to think that this implies a higher potential of error propagation along the flow. This hypothesis is partly supported by [Gangemi *et al.*, 2012], who report a similar drop when they add an automatic disambiguation step followed by an alignment step to DUL classes (including Physical Object). Also for this distinction, SENECA better matches the judgements of the crowd than the experts'.

5.2 Machine Learning Methods

We performed a set of experiments with eight classifiers: J48, Random Forest, REPTree, Naive Bayes, Multinomial Naive Bayes, Support Vector Machines, Logistic Regression, and K-Nearest Neighbors (cf. Section 3.2). We used a 10-fold cross validation strategy using the reference datasets (cf. Section 4). Before training the classifiers, the datasets were adjusted in order to balance the samples of the two classes. The CI_E and PO_E datasets were balanced by randomly removing a set of annotated entities. CI_C and PO_C were balanced by removing entities associated with lower agreement (which constitute weak examples for the classifiers). Each classifier was trained and tested with all four features, described in Section 3.2, both individually and in all possible combinations, with and without performing feature selection. We found that performing feature selection makes the results worse. Having two datasets for each classification (i.e. annotated by the experts and by the crowd) enables multiple configurations of the training set. When we train the classifiers with samples from CI_E and PO_E, they all have the same weight $= 1$. Differently, when the samples come from CI_C and PO_C, they are weighted according to their associated agreement score *agreement(e, c)*, computed with Formula 1 (cf. Section 4). As previously studied by [Aroyo and Welty, 2015], this diverse weighing allows a classifier to learn richer information, including ambiguity and consequent entities that may belong to an "unknown" class, which better represent human cognitive behaviour. Due to space limits, we only report the results of the best performing algorithm[21], which is Support Vector Machine, without feature selection, trained and tested on samples associated with an agreement score $\geq 80\%$. We report on all combinations of features, but **D** alone (i.e. SENECA's output).

Class vs. Instance. Table 4 shows the results of Support Vector Machine, trained on and tested against CI_E and CI_C. The best average performance is obtained with CI_C by combining all features. Combining all features is also the best configuration for each individual classification (i.e. Class (C) and Instance (I)). When CI_E is used there is a slight drop in performance, although the quality of the classification remains high. A possible cause of this result may be the agreement-based weighing provided by CI_C.

Physical Object. Table 4 also shows the results of the Support Vector Machine algorithm trained on and tested against PO_E and PO_C. Similarly to the behaviour of SENECA, statistical approaches worsen their overall performance as compared to the case of the class vs. instance classification. We also observe a different behaviour of the individual features. The best average performance with PO_E is achieved by combining all the feature, while the best average performance with PO_C is achieved by combining the abstract (**A**), the outgoing/incoming properties (**E**), and SENECA output (**D**). In a sense, this confirms that conventions used for creating URI IDs are informative mainly for the class vs. instance distinction.

---

[21]The results of all experiments are available online at https://github.com/fdistinctions/ijcai18

A	U	E	D	Results compared against CI_E							Results compared against CI_C							Results compared against PO_E							Results compared against PO_C						
				P^C	R^C	F_1^C	P^I	R^I	F_1^I	$\overline{F_1}$	P^C	R^C	F_1^C	P^I	R^I	F_1^I	$\overline{F_1}$	P^{PO}	R^{PO}	F_1^{PO}	P^{NPO}	R^{NPO}	F_1^{NPO}	$\overline{F_1}$	P^{PO}	R^{PO}	F_1^{PO}	P^{NPO}	R^{NPO}	F_1^{NPO}	$\overline{F_1}$
✓				.927	.921	.924	.921	.927	.924	.924	.958	.965	.961	.965	.957	.961	.961	.828	.814	.821	.817	.832	.824	.823	.879	.837	.858	.844	.884	.864	.861
	✓			.881	.933	.906	.929	.873	.909	.903	.908	.970	.938	.967	.902	.933	.936	.615	.822	.703	.732	.485	.584	.644	.596	.863	.705	.751	.413	.533	.619
		✓		.854	.975	.911	.971	.834	.897	.904	.886	.983	.932	.981	.874	.924	.928	.786	.865	.824	.857	.764	.805	.814	.782	.886	.831	.868	.752	.806	.818
✓	✓			.928	.935	.932	.935	.928	.931	.932	.966	.971	.968	.971	.966	.968	.968	.831	.829	.838	.833	.832	.831	.831	.851	.811	.831	.819	.857	.838	.834
✓		✓		.939	.943	.941	.943	.939	.941	.941	.971	.976	.974	.976	.971	.973	.974	.869	.867	.868	.867	.870	.868	.868	.912	.853	.882	.862	.918	.889	.885
✓			✓	.934	.927	.939	.928	.934	.931	.931	.966	.964	.965	.964	.966	.965	.965	.849	.829	.835	.831	.843	.837	.836	.865	.834	.849	.839	.869	.854	.852
	✓	✓		.919	.968	.943	.966	.914	.939	.941	.961	.982	.971	.981	.963	.979	.971	.816	.852	.833	.845	.808	.826	.832	.802	**.889**	.842	.875	.777	.823	.833
	✓		✓	.881	.939	.909	.935	.873	.903	.906	.908	.973	.939	.971	.902	.935	.937	.659	.761	.707	.718	.607	.658	.682	.951	.243	.387	.565	**.987**	.719	.553
		✓	✓	.859	.978	.915	**.975**	.846	.903	.909	.889	**.987**	.935	**.985**	.877	.928	.932	.928	.735	.826	.788	.943	.837	.837	**.966**	.762	.852	.803	.973	.886	.866
✓	✓	✓		.942	.946	.944	.945	.942	.944	.944	.973	.981	.976	.980	.973	.976	.976	.865	.866	.866	.866	.865	.865	.865	.927	.847	.885	.859	.933	.894	.891
✓	✓		✓	.939	.933	.936	.934	.939	.937	.936	.968	.969	.968	.969	.968	.968	.968	.831	.824	.828	.826	.833	.829	.828	.878	.831	.855	.838	.876	.856	.853
✓		✓	✓	.945	.949	.943	.941	**.946**	.943	.943	.973	.976	.975	.976	.973	.975	.975	.867	.862	.864	.863	.868	.865	.865	.922	.879	**.899**	**.884**	.923	**.903**	**.901**
	✓	✓	✓	.926	.967	.946	.966	.922	.944	.945	.964	.981	.973	.981	.964	.972	.973	**.933**	.736	.823	.782	**.947**	.857	.843	.962	.759	.849	.801	.975	.877	.863
✓	✓	✓	✓	**.946**	.949	**.947**	.948	**.946**	**.947**	**.947**	**.981**	.982	**.982**	.982	**.981**	**.982**	**.982**	.871	**.877**	**.871**	**.879**	.872	**.871**	**.871**	.905	.866	.886	.872	.909	.895	.888

Table 4: Results of the Support Vector Machine classifier on Class vs. Instance and Physical Object classification task against the reference datasets described in Section 4. The first four columns indicate the features used by the classifier: A is the abstract, U is the URI, E are incoming and outgoing properties, D are the results of the alignment-based methods. P^*, R^*, $\underline{F_1^*}$ indicate precision, recall and F1 measure on Class (C), Instance (I), Physical Object (PO) and the complement of Physical Object (NPO). $\overline{F_1}$ is the average of the F1 measures.

5.3 Remarks and Discussion

We claim that the performed experiments show promising results as far as our research questions are concerned (cf. Section 3). Given the diversity and the basic nature of the two distinctions that we have analysed, and the positive results obtained in both cases by applying the same methods with the same configurations, we claim that the proposed methods can be generalised to other foundational distinctions.

RQ1: Do foundational distinctions match common sense? As anticipated in Section 4 the high agreement observed among workers that participated in the crowdsourcing tasks, as well as the high agreement between the crowd and the experts, suggest that the foundational distinctions that we have tested do actually match common sense.

RQ2: Does the (Semantic) Web provide an empirical basis for supporting foundational distinctions over LOD entities, according to common sense? We claim that the high average value of F1 measure associated with all experiments indicates that the Web, and in particular LOD, implicitly encodes foundational distinctions. We also think that, more in general, this is a hint that the Web is a good source for common sense knowledge extraction. We find particularly interesting to observe that the feature **E** (i.e. ongoing/incoming properties) has always a positive impact, in all features combinations, on the classifier's performance (cf. Table 4), for both tasks. This motivates us in conducting further investigations (i) towards identifying and testing additional features based on LOD, e.g. more sophisticated use of assertions and axioms from LOD as well as (ii) to analyse LOD at a much larger scale (e.g. by using LOD Laundromat [Beek *et al.*, 2016]) with an *empirical science* perspective: looking for emerging patterns that may encode relevant pieces of common sense knowledge [Gangemi and Presutti, 2010]. Our promising results open a number of possible research directions: besides replicating these experiments at a larger scale, we plan a follow up study concerning the application of the same approach to distinguishing physical objects that can act as locations for other physical objects. This is particularly relevant in order to extract knowledge about where things are usually located in, whether a location is appropriate for an object in terms of its size, etc. Another relevant distinction to be investigated with priority is the one between physical and social objects (e.g. organisations), which

is often prone to *systematic polysemy* [Pustejovsky, 1998], i.e. objects that have a same linguistic reference, but different (disjoint) types of referents. For example, the term *National Library* is used to refer both to an organisation (a social object) taking care of the library's collections, and of the related administrative and organisational issues, and to the buildings (physical objects) where the organisational staff works and the collections are located in. Besides covering foundational distinctions, we aim to extend our approach to learn or discover relational knowledge such as the one modelled and encoded in terms of frames [Fillmore, 1982; Gangemi *et al.*, 2016].

RQ3: What ensemble of features, resources, and methods works best to make machines learn foundational distinctions over LOD entities? According to our results, statistical methods perform better than alignment-based methods. We use supervised learning and crowdsourcing to test two very diverse foundational distinctions, both very basic in knowledge representation and foundational ontologies. It emerges that two features show the same ability to positively impact on the methods' performance, for both distinctions: **A** (a text describing the entity) and **E** (entity's outgoing and incoming properties). Both features convey the semantic description of an entity: the former by means of natural language, which characterises a huge portion of the web, the latter by means of LOD triples, which characterise the semantic web. Based on these observations, we argue that the method can be generalised, even if each specific distinction may benefit from a specialised extension of the feature set. In our case, the **U** feature (i.e. URI ID) clearly shows effectiveness for the class vs. instance rather than for PO vs. NPO.

A question is whether DBpedia text is special because of its "standardised" style of writing. Our experiments and results do not cover this issue, which needs to be assessed in order to provide a stronger support to our claim of generalisability. A similar doubt can be raised as far as outgoing and incoming links are concerned. DBpedia properties mainly come from infoboxes, which also follow, and are influenced by the standardised way of writing Wikipedia pages. Nevertheless, for this feature we argue that the doubt does not apply, since incoming and outgoing properties include links to and from LOD datasets that are outside DBpedia, hence independent from the standardised content of Wikipedia.

6 Conclusion

This study reports a set of experiments for assessing whether the Web, and in particular Linked Open Data, provides an empirical basis to extract foundational distinctions, and if they match common sense. For testing the former, we adopt and compare two approaches, namely alignment-based methods and machine learning methods. For the latter we use crowdsourcing and compare the judgements of the crowd with those of experts'. For both questions we observe promising results and define a method that can be generalised to investigate additional distinctions. We plan experiments on other foundational distinctions (e.g. types of locations, objects that can serve as locations or containers, etc.) and with additional methods. Our ultimate goal is to advance the state of the art of AI tasks requiring common sense reasoning by designing a methodological framework that enables mass-production of common sense knowledge, and its injection into LOD.

Acknowledgements

References

[Aroyo and Welty, 2015] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[Beek *et al.*, 2016] Wouter Beek, Laurens Rietveld, Stefan Schlobach, and Frank van Harmelen. LOD laundromat: Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing*, 20(2):78–81, 2016.

[Bizer *et al.*, 2009] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *International Journal of Web Semantics*, 7(3):154–165, 2009.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36 – 64, 2016.

[Davis and Marcus, 2015] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[Fellbaum, 1998] Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

[Fillmore, 1982] Charles J. Fillmore. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. 1982.

[Gangemi and Presutti, 2010] Aldo Gangemi and Valentina Presutti. A pattern science for the semantic web. *Semantic Web*, 1(1-2):61–68, 2010.

[Gangemi *et al.*, 2003] Aldo Gangemi, Roberto Navigli, and Paola Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In *Proceedings of COOPIS, DOA and ODBASE Confederated International Conferences*, pages 3–7, 2003.

[Gangemi *et al.*, 2012] Aldo Gangemi, Andrea Giovanni Nuzzolese, Valentina Presutti, Francesco Draicchio, Alberto Musetti, and Paolo Ciancarini. Automatic typing of dbpedia entities. In *Proceedings of the International Semantic Web Conference, Part I*, pages 65–81, 2012.

[Gangemi *et al.*, 2016] Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. Framester: A wide coverage linguistic linked data hub. In *Proceedings of Knowledge Engineering and Knowledge Management*, pages 239–254, 2016.

[Hayes, 1989] Patrick Hayes. The second naive physics manifesto. In *Readings in qualitative reasoning about physical systems*, pages 567–585. 1989.

[Lakoff, 1987] George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, 1987.

[Lenat, 1995] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):32–38, 1995.

[Liu and Singh, 2004] Hugo Liu and Push Singh. ConceptNet –a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[Masolo *et al.*, 2003] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. WonderWeb Deliverable D18 - Ontology Library (Final Version), 2003.

[Miller and Hristea, 2006] George A. Miller and Florentina Hristea. Wordnet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3, 2006.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012.

[Paulheim and Gangemi, 2015] Heiko Paulheim and Aldo Gangemi. Serving dbpedia with DOLCE - more than just adding a cherry on top. In *Proceedings of the International Semantic Web Conference, Part I*, pages 180–196, 2015.

[Pease and Niles, 2002] Adam Pease and Ian Niles. IEEE standard upper ontology: a progress report. *The Knowledge Engineering Review*, 17(1):65–70, 2002.

[Pustejovsky, 1998] James Pustejovsky. *The Generative Lexicon*. MIT Press, 1998.

[Russell and Norvig, 2009] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, 3rd edition, 2009.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, 2007.