Distributional semantics for ontology verification *

Julien Corman
IRIT, University of Toulouse
julien.corman@irit.fr

Nathalie Aussenac-Gilles
CNRS
IRIT, University of Toulouse
aussenac@irit.fr

Laure Vieu
CNRS
IRIT, University of Toulouse
vieu@irit.fr

Abstract

As they grow in size, OWL ontologies tend to comprise intuitively incompatible statements, even when they remain logically consistent. This is true in particular of lightweight ontologies, especially the ones which aggregate knowledge from different sources. The article investigates how distributional semantics can help detect and repair violation of common sense in consistent ontologies, based on the identification of consequences which are unlikely to hold if the rest of the ontology does. A score evaluating the plausibility for a consequence to hold with regard to distributional evidence is defined, as well as several methods in order to decide which statements should be preferably amended or discarded. A conclusive evaluation is also provided, which consists in extending an input ontology with randomly generated statements, before trying to discard them automatically.

1 Introduction

Ontology learning from texts deals with the automated extraction of knowledge from linguistic evidence. This article investigates a slightly different problem, which is how Natural Language Processing may provide hints for the identification of statements of an input ontology which are unlikely to hold if the rest of it does. As a minimal example, consider the following set Δ of statements, from DBpedia (Mendes et al., 2012), and assume that Δ is a subset of a larger set of statements K (for instance DBpedia itself, or some subset of it) :

Ex 1.
$\Delta = \{$(1) `keyPerson`(*Caixa Bank*, *CEO*),
(2) `keyPerson`(*BrookField Office Properties*, *Peter Munk*)
(3) `occupation`(*Peter Munk*, *CEO*) $\}$

There is a clear violation of common sense in Δ : the individual *CEO* must be both a key person of *Caixa Bank*, and the occupation of another individual (*Peter Munk*), who is himself a key person of some company. Detecting such cases within (larger) sets of logical statements is of particular interest in OWL, which facilitates the aggregation of knowledge from multiple sources with overlapping signatures, yielding datasets in which several incompatible understandings of a same individual or predicate may coexist. This easily leads to undesired inferences, even when the dataset is logically consistent.[1] But as the example illustrates, the problem may also occur within a single knowledge base, especially if it has been built semi-automatically, and/or is issued from a collaborative effort.

Another problem of interest consists in deciding which statement(s) should be preferably discarded or amended in order to get rid of the nonsense. In example 1, without further information, it would be intuitively relevant to discard or modify either (1) or (2). Unfortunately though, Δ alone does not give any indication of which of the two should be preferably discarded. But the whole input ontology $K \supset$

* The research reported here was supported by a Marie Curie FP7 Career Integration Grant, Grant Agreement Number PCIG13-GA-2013-618550.

[1] and *coherent* in the Description Logics sense, i.e. whose signature contains unsatisfiable DL atomic concepts/OWL named classes

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 30–39,
Denver, Colorado, June 4–5, 2015.

Δ may. To keep the example simple, let us assume that *Peter Munk*, *CEO* and `occupation` do not appear in $K \setminus \Delta$. Then a reasonable assumption is that the overall understanding of `keyPerson` within K should be the decisive factor. If it generally ranges over person functions (i.e. if in most instances of the relation according to K, the second argument is a person function), then it is to be understood as "has as a key person someone whose function is", and (2) should be preferably discarded. Alternatively, if `keyPerson` generally ranges over human beings, then (1) should be preferably discarded.

The article investigates the use of linguistic evidence to solve both of these problems : identifying violations of common sense, and selecting the statement(s) to be preferably amended or discarded. This may be viewed as a small paradigm shift, in that it questions an assumption commonly made in the knowledge extraction literature, namely that manually crafted knowledge strictly prevails over the one obtained from linguistic sources. By default, the case of a consistent[2] input ontology K will be studied, but section 6 discusses the application of the approach to an inconsistent K as well.

As a concrete contribution, section 5 evaluates the adaptation of relatively simple techniques issued from named entity classification/ontology population, and based on distributional semantics. To illustrate how this works, let us assume that the only other appearance of `keyPerson` within K is the following OWL statement :

(4) `hasRange(keyPerson, Person)`

i.e. in FOL :

(4) $\forall xy(\texttt{keyPerson}(x, y) \rightarrow \texttt{Person}(y))$

Then $K \models \psi_1 = \texttt{Person}(\textit{CEO})$, and $K \models \psi_2 = \texttt{Person}(\textit{Peter Munk})$. Assume also that there are other instances of `Person` according to K, and that most of them are actually human beings (like *Peter Munk*). Then ψ_1 is an undesirable consequence of K, whereas ψ_2 on the other hand reinforces it.

Distributional semantics characterizes a word (or possibly a multi word unit) by some algebraic representation of the linguistic contexts with which it is observed. These representations have already been used for ontology population, for instance by (Tanev and Magnini, 2008), the main intuition being that individuals denoted by linguistic terms with similar contexts tend to instantiate the same classes. The underlying linguistic phenomenon is known as *selectional preference*, i.e. the fact that some contexts tend to select or rule out certain categories of individuals : e.g. the context "*X* was born in" tends to select a human being, whereas "*X* was launched" tends to rule it out. Back to the example, one can expect the similarity between the distributional representation of the term "C.E.O" and other terms denoting instances of `Person` according to K to be relatively low, hindering the plausibility of ψ_1 with regard to K. In other words, ψ_1 should stand as an outlier among consequences of K, and therefore is probably undesirable. Conversely, the similarity between "Peter Munk" and terms denoting other instances of `Person` should be relatively high. For simplicity, suppose that (1), (2), (3) and (4) are the only 4 statements of K which are candidate for removal. Then in order to give up the belief in ψ_1 while preserving ψ_2, it is necessary to discard (1), and retain (2) and (4). It is also sufficient to discard (1), i.e. discarding (3) as well would result in an unnecessary information loss. So in this case, the evidence provided by distributional semantics should suggest the removal of (1), or at least its modification, which is also intuitively the correct solution.

Section 4 formalizes this approach, by defining a score which estimates the plausibility of some consequences a subbases Γ of K, given distributional evidence. Section 5 then provides an original evaluation of this strategy, based on the prior extension of a small OWL ontology with randomly generated statements. The approach is evaluated for both problems, i.e. the identification of undesired consequences and statements. Performances of several forms of distributional representations are also compared. Section 6 discusses immediate applications, in particular for (consistent and inconsistent) ontology debugging. Finally, section 7 considers possible extensions of this framework, as well as their limitations. Section 2 is a brief overview of related works in the fields of ontology learning and debugging, whereas section 3 introduces notational conventions, and lists some preliminary requirements to be met by the input K.

[2] and coherent (see footnote 1)

2 State of the art

Ontology learning from texts (Cimiano, 2006; Buitelaar et al., 2005) aims to automatically build or enriching a set of logical statements out of linguistic evidence, and is closely related to the field of information extraction. The work presented here borrows from a subtask called ontology population (which itself borrows from named entity classification), but only when the individuals and concepts of interest are already known (Cimiano and Völker, 2005; Tanev and Magnini, 2008; Giuliano and Gliozzo, 2008), which is not standard. A comparison may also be drawn with the use of linguistic evidence by (Suchanek et al., 2009) for information extraction in the presence of conflicting data.

But the objective of the present work is different, pertaining to ontology debugging, which covers a wide range of techniques, from syntactic verifications (Poveda-Villalón et al., 2012) to anti-patterns detection (Roussey and Zamazal, 2013), both based on common modeling mistakes, or the submission of models (Ferré and Rudolph, 2012; Benevides et al., 2010) or consequences (Pammer, 2010) of the input ontology to the user. As discussed in section 6, the framework depicted here presents an interesting complementarity with debugging techniques developed in the Description Logics community, prototypically based on diagnosis (Friedrich and Shchekotykhin, 2005; Kalyanpur et al., 2006; Qi et al., 2008; Ribeiro and Wassermann, 2009), because they require the prior identification of some undesired consequence of K (be it $\bot$). But distributional evidence may also provide a principled way of selecting most relevant diagnoses among a potentially large number of candidates, as well as an alternative to their exhaustive computation, which has been shown costly by (Schlobach, 2005).

3 Conventions and presuppositions

The prototypical input is a set of statements in OWL DL or OWL 2, although the approach may be generalized to other representation languages. OWL DL and OWL 2 are based on Description Logics (DL), which are themselves decidable fragments of first-order logic (FOL). The OWL notation is preferred to the DL one for readability, and FOL translations are given when not obvious.

An *ontology* is just understood here as a (finite) set of logical statements. A *class* will designate a named class in OWL, i.e. a FOL unary predicate, like `Person`, whereas a *named individual*, or just *individual*, designates a constant, like *Peter Munk*.

The input ontology K must provide English terms denoting some of its named individuals (e.g. the term "Peter Munk"). These terms are prototypically named entities, but may also occasionally be common nouns (or common noun phrases), as shown in example 1 with "C.E.O". There may be multiple terms for a same individual. The approach cannot handle polysemy though, in particular the fact that some individuals of K may have homonyms (within K or not), for instance that the term "JFK" can stand for a politician, airport or movie. Ideally, no distributional representation should be built for individuals of K with potential homonyms. Some of them may be identified with simple strategies, like checking the existence of a Wikipedia disambiguation page. On the opposite, labels for classes of K (prototypically common nouns or common noun phrases, which are arguably more ambiguous) are never used during the process.

4 Proposition

Given a subbase Γ of the input ontology K (possibly K itself), the ontology verification strategy presented in introduction relies on the evaluation of a set Ψ_Γ of consequences of Γ. This section first defines a score $\mathrm{sc}_\Gamma(\psi)$ for each $\psi \in \Psi_\Gamma$, which intuitively evaluates the plausibility of ψ wrt Γ, provided some distributional representation for each named individual appearing in Ψ_Γ. Then it discusses how this score can be used to select statements of the input ontology K which, according to distributional evidence, should be preferably discarded, or at least amended.

4.1 Plausibility of a consequence $\psi \in \Psi_\Gamma$

For the experiments described in section 5, Ψ_Γ is the set of consequences of Γ of the form $A(e)$ or $\neg A(e)$, with e a constant (like *CEO*) and A a unary predicate (like `Person`), and for which linguistic occurrences of a term denoting e could be retrieved. Possible extension of Ψ_Γ with other types of formulas is discussed in section 7.

Let ψ be a formula of Ψ_Γ, of the form $A(e)$, e.g. $\psi = \texttt{Person}(\textit{CEO})$ or $\psi = \texttt{Person}(\textit{Peter Munk})$. Then $\text{inst}_\Gamma(A)$ will designate all instances of A according to Γ for which linguistic occurrences could be retrieved, i.e. $\text{inst}_\Gamma(A) = \{e' \mid A(e') \in \Psi_\Gamma\}$, and $\text{inst}_\Gamma(A) \setminus \{e\}$ will be called the *support set* for $A(e)$. Similarly, $\text{inst}_\Gamma(\top)$ will designate all named individuals appearing in Ψ_Γ.

Let $\text{sim}(e_1, e_2)$ be a measure of similarity between the distributional representations of individuals e_1 and e_2 (prototypically the cosine similarity between some vector representations of the linguistic contexts of e_1 and e_2). Then for each $e' \in \text{inst}_\Gamma(A) \setminus \{e\}$, if $\text{sim}(e, e')$ is lower than what could be expected if e' was a random individual of $\text{inst}_\Gamma(\top) \setminus \{e\}$ (i.e. not necessarily an instance of A), the hypothesis that $A(e)$ is an outlier within Ψ_Γ will be reinforced.

For instance, in example 1, let $\psi = \texttt{Person}(\textit{CEO})$ and $\Gamma = K$. Then the support set $\text{inst}_\Gamma(A) \setminus \{e\}$ is composed of all other instances of `Person` according to Γ. For each individual e' of this support set, if $\text{sim}(\textit{CEO}, e')$ is lower than what can be expected for a random individual of K with linguistic occurrences (and different from *CEO*), then the confidence in $\texttt{Person}(\textit{CEO})$ should decline. Conversely, if $\text{sim}(e, e')$ is higher that expected, the hypothesis that ψ is in line with Ψ_Γ will be reinforced.

Here is a cost-efficient and relatively simple method to compute a plausibility score $\text{sc}_\Gamma(A(e))$. Let $S = \text{inst}_\Gamma(A) \setminus \{e\}$ designate the support set for Γ and e, and $|S|$ the cardinality of S, i.e. the number of other instances of A according to Γ. And let us assume a set W of $|S|$ randomly chosen elements of $\text{inst}_\Gamma(T) \setminus \{e\}$, i.e. of $|S|$ individuals which are different from e, but not necessarily instances of A. Finally, let the random variable $X^\Gamma_{e,|S|}$ model the expected value of $\sum_{e' \in W} \frac{\text{sim}(e,e')}{|S|}$, i.e. the mean of the similarities between e and each individual of W. In other words, if $|S|$ individuals were randomly chosen instead of those of the support set, $X^\Gamma_{e,|S|}$ models what the average similarity between e and these individuals can be expected to be. Then the plausibility $\text{sc}_\Gamma(A(e))$ of $A(e)$ can be defined by :

Definition 4.1. If $S = \text{inst}_\Gamma(A) \setminus \{e\}$, then $\text{sc}_\Gamma(A(e)) = p(X^\Gamma_{e,|S|} \leq \sum_{e' \in S} \frac{\text{sim}(e,e')}{|S|})$

$\text{sc}_\Gamma(A(e))$ estimates of how surprisingly high the similarity between e and the individuals of the support set S is, considering the overall similarity between e and the individuals of Γ.

For the evaluation described in section 5, the random variable $X^\Gamma_{e,|S|}$ was assumed to follow a beta distribution $\text{Beta}(\alpha, \beta)$, which intuitively allows taking the size $|S|$ of the support set into account. For instance, if $S = \{e'\}$, i.e. $|S| = 1$, then *ceteris paribus* a high similarity between e and e' will be less informative than an equally high average similarity between e and all elements of a large S. Stated another way, the lower $|S|$ is, the more uniform the distribution of $X^\Gamma_{e,|S|}$ should be. This can be obtained by setting $X^\Gamma_{e,|S|} \sim \text{Beta}(m|S| + 1, (1 - m)|S| + 1)$, where m is the average similarity between e and all other individuals of the signature of Γ, i.e. $m = \sum_{e' \in \Gamma \setminus \{e\}} \frac{\text{sim}(e,e')}{|\Gamma|-1}$.

A possible interrogation here is the choice of $\text{inst}_\Gamma(A) \setminus \{e\}$ as the support set for $A(e)$. For instances, if $\psi = \texttt{Person}(\textit{Peter Munk})$, a case could be made for using $\text{inst}_\Gamma(\neg A)$ as well, i.e. for exploiting the (dis)similarity between *Peter Munk* and individuals which, according to K, are instances of $\neg\texttt{Person}$.[3] This is quite unrealistic though from a linguistic point of view, which can be intuitively seen in this example by replacing *Peter Munk* with *CEO*. Assume for instance that *Thelonious Monk* and *Beijing* are (reliable) instances of `Person` and $\neg\texttt{Person}$ respectively according to Γ. There is no reason to expect that $\text{sim}(\textit{CEO}, \textit{Beijing}) > \text{sim}(\textit{CEO}, \textit{Thelonious Monk})$. In other words, it is implausible to assume that elements of $\text{inst}_\Gamma(\neg A)$ should *a priori* share similar contexts.

Interestingly enough, and for the same reason, the support set for a consequence of Γ of the form $\neg A(e)$ is not $\text{inst}_\Gamma(\neg A)$, but $\text{inst}_\Gamma(A)$, which yields :

Definition 4.2. If $S = \text{inst}_\Gamma(A)$, then $\text{sc}_\Gamma(\neg A(e)) = p(X^\Gamma_{e,|S|} \geq \sum_{e' \in S} \frac{\text{sim}(e,e')}{|S|})$

[3] i.e. $\Gamma \models \neg\texttt{Person}(e')$ not only $\Gamma \not\models \texttt{Person}(e')$

4.2 Linguistic compliance of Γ

This does not directly address the second problem mentioned in introduction though. For practical ontology verification, it is also desirable to identify the cause of this nonsense, i.e. statements (*axioms* in the DL terminology) which are intuitively problematic. For instance, in example 1, computing $\mathrm{sc}_\Gamma(\psi)$ for each $\psi \in \Psi_K$ may signal that the consequence ψ_1 is unlikely to hold wrt the larger ontology K. And discarding either (1) or (4) is sufficient to get rid of the belief in ψ. But given the additional assumptions made about K, discarding the former is preferable, in that discarding the latter would also result in the loss of ψ_2. In other words, some subbases of K (like $K \setminus (1)$ here) are more relevant than others (e.g. $K \setminus (4)$), which can be simply captured as follows. Let $\mathrm{comp}(\Gamma)$ be an estimation of the compliance of a subbase Γ of K with the gathered linguistic evidence. A straightforward option consists in setting $\mathrm{comp}(\Gamma)$ to be the mean of the scores of evaluated consequences for Γ, i.e. :

Definition 4.3. $\mathrm{comp}(\Gamma) = \sum_{\psi \in \Psi_\Gamma} \frac{\mathrm{sc}_\Gamma(\psi)}{|\Psi_\Gamma|}$

Then a strict partial order $\prec$ over 2^K can simply be defined by $\Gamma_1 \prec \Gamma_2$ iff either $\mathrm{comp}(\Gamma_1) < \mathrm{comp}(\Gamma_2)$, or ($\mathrm{comp}(\Gamma_1) = \mathrm{comp}(\Gamma_2)$ and $\Gamma_1 \subset \Gamma_2$),[4] and a subbase Γ of K can be viewed as optimal if it is maximal wrt $\prec$.[5]

In practice though, identifying optimal subbases is a non trivial task. To see this, note that the function to be maximized is not directly a function of the statements in Γ, but of Ψ_Γ, i.e. some of the consequences of Γ. So even if one could identify a subset Ψ' of Ψ_K which maximizes this function, there may not exist a subbase Γ of K such that $\Psi_\Gamma = \Psi'$. Another difficulty comes from the fact that for two subbases Γ_1 and Γ_2 of K, and a consequence $\psi \in \Psi_{\Gamma_1} \cap \Psi_{\Gamma_2}$, it doesn't hold in general that $\mathrm{sc}_{\Gamma_1}(\psi) = \mathrm{sc}_{\Gamma_2}(\psi)$, because the support set for ψ in Γ_1 may differ from its support set in Γ_2. In particular, it may be the case that $\Gamma_1 \subseteq \Gamma_2$ but $\mathrm{sc}_{\Gamma_1}(\psi) > \mathrm{sc}_{\Gamma_2}(\psi)$, which greatly reduces the possible uses of monotonicity (if $\Gamma_1 \subseteq \Gamma_2$, then $\mathrm{Cn}(\Gamma_1) \subseteq \mathrm{Cn}(\Gamma_2)$) to optimize the exploration of 2^K. More generally, if the optimal subbases of K are small (say twice smaller that K), it can be rightfully argued that dropping so many statements for the sake of linguistic evidence is not a viable debugging strategy.

Therefore a more plausible application scenario is one in which the search space has been previously circumscribed, either by setting a maximal (small) number of statements to discard, or by identifying a set of potentially erroneous statements, through *axiom pinpointing*, as explained in section 6. This is also why the evaluation presented in section 5 focuses on the simplest possible case, i.e. the removal from K of one statement only, whereas the integration of distributional evidence to more complex debugging strategies is discussed in section 6.

As an alternative to the function comp, and in order to avoid the fact that a same consequence may have different plausibility scores wrt two subbases of K, one may choose to discard unlikely consequences based on their respective scores in K, i.e. to use the score $\mathrm{comp}_K(\Gamma)$,[6] defined by :

Definition 4.4. $\mathrm{comp}_K(\Gamma) = \sum_{\psi \in \Psi_\Gamma} \frac{\mathrm{sc}_K(\psi)}{|\Psi_\Gamma|}$

This solution is arguably less satisfying, but more amenable to optimizations. A trivial example is that of a subbase Γ_1 with $\max_{\psi \in \Psi_{\Gamma_1}} \mathrm{sc}_K(\psi) < \mathrm{comp}_K(\Gamma_2)$ for some already evaluated subbase Γ_2, in which case no subbase of Γ_1 can be optimal wrt $\prec$.

Additionally, instead of taking the mean of the scores of evaluated consequences of Γ, one may want to penalize the subbases of K with the most unlikely consequences, which gives a standard (total) lexicographic ordering $\preceq_{lex}$ on 2^K, defined as follows. Let $\boldsymbol{\omega}_\Gamma = \omega_\Gamma^1, .., \omega_\Gamma^{|\Psi_\Gamma|}$ be the vector of formulas of Ψ_Γ order by increasing score sc_Γ, and let $\mathrm{sc}_\Gamma(\boldsymbol{\omega}_\Gamma) = \mathrm{sc}_\Gamma(\omega_\Gamma^1), .., \mathrm{sc}_\Gamma(\omega_\Gamma^{|\Psi_\Gamma|})$. Then $\preceq_{lex}$ is defined by $\Gamma_1 \preceq_{lex} \Gamma_2$ iff either $\mathrm{sc}_{\Gamma_1}(\boldsymbol{\omega}_{\Gamma_1}) = \mathrm{sc}_{\Gamma_2}(\boldsymbol{\omega}_{\Gamma_2})$, or (there is a $1 \leq i \leq |\Psi_{\Gamma_2}|$ such that $\mathrm{sc}_{\Gamma_1}(\omega_{\Gamma_1}^j) = \mathrm{sc}_{\Gamma_2}(\omega_{\Gamma_2}^j)$ for all $1 \leq j < i$, and either $\mathrm{sc}_{\Gamma_1}(\omega_{\Gamma_1}^i) < \mathrm{sc}_{\Gamma_2}(\omega_{\Gamma_2}^i)$ or $|\Psi_{\Gamma_1}| = i - 1$). Then

[4] The assumption is made that a minimum of syntactic information should be lost whenever possible, i.e. Γ_1 and Γ_2 are primarily viewed as bases, not as theories. In particular, if $\mathrm{Cn}(\Gamma_1) = \mathrm{Cn}(\Gamma_2)$, but $\Gamma_1 \not\subseteq \Gamma_2$ and $\Gamma_2 \not\subseteq \Gamma_1$, then Γ_1 and Γ_2 are not comparable wrt $\prec$. Redundancies in this view should also be preserved when possible, i.e. if $\mathrm{Cn}(\Gamma_1) = \mathrm{Cn}(\Gamma_2)$ and $\Gamma_1 \subset \Gamma_2$, then $\Gamma_1 \prec \Gamma_2$ still holds.

[5] There may be several several optimal subbases.

[6] or more generally $\mathrm{comp}_{\Gamma'}(\Gamma)$, for some $\Gamma' \supseteq \Gamma$

as previously, a strict partial order $\prec$ over 2^K can be defined by $\Gamma_1 \prec \Gamma_2$ iff either $\Gamma_1 \prec_{lex} \Gamma_2$, or ($\Gamma_1 =_{lex} \Gamma_2$ and $\Gamma_1 \subset \Gamma_2$).

Again, $\mathrm{sc}_K(\psi)$ may be used instead of $\mathrm{sc}_\Gamma(\psi)$, yielding the lexical ordering $\preceq_{lex_K}$. This last possibility corresponds to a relatively intuitive operation, which consists in giving up in priority the most implausible consequences of K. All four possibilities are evaluated in what follows.

5 Evaluation

The dataset used for this evaluation is a fragment of the fisheries ontology from the NEON project.[7] It has been automatically built out of 10 randomly selected named individuals, applying a module extraction procedure, followed by a trimming algorithm. The fragment contains 1038 (logical) statements, and involves 71 named individuals (mostly geographical or administrative entities), the least expressive underlying DL being $\mathcal{SI}$.

The linguistic input is a small corpus of approximately 6300 web pages, retrieved with a search engine, using the labels of named individuals of F as queries. The HTML documents were cleaned with the BootCat library (Baroni and Bernardini, 2004).

The construction of the distributional representations of the named individuals of F was basic, the use of more elaborate methods (SVD,...) being left for future work. The approach presented in this article remains generic enough to be applied to most existing distributional frameworks, the only requirement being a real-valued similarity measure.

Two different forms of linguistic contexts were alternatively tested. The first option considers as a context any n-gram ($2 \leq n \leq 5$) without punctuation mark which immediately precedes or follows a term t denoting an individual of F. The other option is a more customized one, extracting sequences of lemmatized words (*lemmaPOS* in what follows) surrounding t, in a shifting window of 3 to 5 tokens + the size of t, ignoring certain categories of word. Part-of-speech tagging was performed thanks to the Stanford Parser (Toutanova et al., 2003), with a pre-trained model for English. If *Cont* designates the set of contexts observed with at least 2 individuals, then an individual was represented by the vector of its respective frequencies with each context $c \in$ *Cont*. Different possibilities were compared to weight these frequencies. The pointwise mutual information (PMI) was used in a standard way for n-grams and lemmaPOS contexts (with possible negative resulting frequencies set to 0). Following (Giuliano and Gliozzo, 2008), the self-information *self*(c) was also used for n-grams, defined by *self*$(c) = -\log p(c)$, the probability $p(c)$ being estimated thanks to the Microsoft Web N-gram Services. A combined weighting by PMI and self-information was also tested for n-grams. These alternative settings are represented by capital letters in tables 1 and 2 : LP for lemmaPOS with PMI, and NP, NS and NPS for n-grams with PMI, self-information and both respectively.

The ontology F has been extended for the sake of the evaluation, with statements randomly generated out of its signature. The underlying assumption is that adding such statements to F is very likely to generate violations of common sense (although nothing prevents in theory the generation of plausible statements too). The goal for the evaluation was then to automatically retrieve proper consequences of each extension of F on the one hand, and the random statements themselves on the other hand.

To prevent any misunderstanding, it should be emphasized that this is not a realistic application case. The input ontology was selected for its quality, and degraded through random statement generation, allowing an arguably artificial, but also very objective evaluation procedure (the only bias may come from randomly generated statements which are actually plausible). By contrast, using a non modified input dataset, and evaluating whether or not the axioms/consequences spotted by the algorithm are actually erroneous is a complex and subjective task, with a possibly low inter-annotator agreement.

The generation procedure randomly selects a statement $\phi \in F$, and yields a statement ϕ' with the same syntactic structure as ϕ, but in which individuals and predicates have been replaced by random individuals and predicates appearing in F. For instance, if $\phi = \forall xy(A(x) \wedge r(x,y) \rightarrow \neg B(y))$, then $\phi' = \forall xy(C(x) \wedge s(x,y) \rightarrow \neg D(y))$, with C and D (resp. s) randomly chosen among classes (resp. binary predicates) of the signature of F.

100 randomly generated statements $\phi_1, \ldots, \phi_{100}$

[7] http://www.neon-project.org/nw/Ontologies

	rank	p-val
LP	4.15 / 216.1	<0.001
NP	9.73 / 216.1	<0.001
NS	7.33 / 216.1	<0.001
NPS	5.59 / 216.1	<0.001

Table 1: Average ranking among Ψ_{K_i} of the lowest-ranked formula of $\Psi_{K_i}^{rand}$, and p-value for the rankings of all formulas of all $\Psi_{K_i}^{rand}$

were added independently to F, yielding 100 input ontologies $K_1, \ldots, K_{100}$, such that each K_i was consistent, and that there was at least one consequence of the form $A(e)$ or $\neg A(e)$ entailed by K_i but not by F, with e sharing at least one linguistic context with some other individual of F. All 100 input ontologies are available online.[8]

The first part of the evaluation was performed as follows. For each K_i and each $\psi \in \Psi_{K_i}$, the plausibility $\mathrm{sc}_{K_i}(\psi)$ was computed as in definitions 4.1/4.2, and Ψ_{K_i} was ordered by increasing plausibility.[9] Within Ψ_{K_i} are consequences which were not initially entailed by F, but have been obtained after the extension of F with the random statement ϕ_i. So in a sense, these consequences are randomly generated too, and therefore one may expect many of them to convey absurd information (for instance `Architect`(***Belgium***)), or at least to be outliers (like `Person`(***CEO***) in ex 1) within Ψ_{K_i}. Let $\Psi_{K_i}^{rand}$ designate these additional consequences, i.e. $\Psi_{K_i}^{rand} = \Psi_{K_i} \setminus \Psi_F$. If $\psi \in \Psi_{K_i}^{rand}$, and if $\mathrm{sc}_{K_i}(\psi)$ is actually lower than for most other formulas of Ψ_{K_i}, this would indicate that the plausibility score, as formulated in definitions 4.1/4.2, is actually a good estimator.

In order to evaluate this, column "rank" in table 1 gives the average ranking (for all 100 ontologies) within Ψ_{K_i} of the formula $\psi_i \in \Psi_{K_i}^{rand}$ with lowest score. The lower this ranking, the more efficient the plausibility score is at detecting outlier consequences. Column "pVal" gives the probability (t-test) for the cumulated rankings of all formulas in all $\Psi_{K_i}^{rand}$ to be as low as the observed ones, if all consequences in all Ψ_{K_i} had been randomly ordered.

[8] http://www.irit.fr/~Julien.Corman/index_en.php

[9] The ranking was a strict ordering : if two consequences had the same score, one of them was randomly designated as strictly lower ranked.

Results are convincing, with a significant p-value for all four settings. For most ontologies (75/100), there was only one formula in $\Psi_{K_i}^{rand}$. A closer look at the data revealed that, for the best setting (LP), in most of theses cases (57/75), the only formula in $\Psi_{K_i}^{rand}$ was also the one with lowest plausibility in Ψ_{K_i}, over 216.1 on average, i.e. the only randomly generated consequence was also the least plausible one according to linguistic evidence. This is very encouraging, especially considering the relatively small number of named individuals (71) in F, i.e. the fact that the support to evaluate the plausibility of a consequence $\psi \in \Psi_{K_i}$ was limited. On the other hand, performances were generally poor when the cardinality of $\Psi_{K_i}^{rand}$ was important ($> 0.25 * |\Psi_{K_i}|$), which may be explained by the fact that support sets for some classes of F were significantly modified after the extension of F with ϕ_i.

As for the settings, unsurprisingly, the two most beneficial (but unfortunately incompatible) factors were the use of lemmatized contexts on the one hand (LP), and the queries over the Web N-gram corpus on the other hand (NS and NPS)

The second part of the evaluation focused on the retrieval of the random statements $\phi_1, .., \phi_{100}$, for the LP setting only, because it gave the best results in the previous experiment. For each extended base K_i, all immediate subbases $\Gamma_{i,1}, .., \Gamma_{i,|F|+1}$ of K_i were generated, i.e. each $\Gamma_{i,j}$ was such that $K_i = \Gamma_{i,j} \cup \{\phi_j\}$ for some statement ϕ_j of K_i. The different $\Gamma_{i,j}$ were ordered by decreasing compliance score $\mathrm{comp}(\Gamma_{i,j})$ (resp. $\mathrm{comp}_{K_i}(\Gamma_{i,j})$), or by decreasing lexicographic ordering $\preceq_{lex}$ (resp. $\preceq_{lex_{K_i}}$).[10] Intuitively, this yields a ranking on K_i where the least reliable statements wrt linguistic evidence should appear first : if $\phi_j \in K_i$, and if the subbase of K_i obtained by discarding ϕ_j (i.e. $\Gamma_{i,j}$) has a higher linguistic compliance score than K_i, then discarding $\Gamma_{i,j}$ can be viewed as an improvement over K_i. And if $\Gamma_{i,j}$ is among the best ranked subbases of K_i, then ϕ_j is among the least reliable statements of K_i wrt distributional evidence. For instance, in example 1, one may expect the subbase $K \setminus (1)$ to have a maximal linguistic compliance score among immediate subbases of K (or to be

[10] Again, the ranking was randomly turned into a strict ordering (see footnote 9).

	rank	p-val
$\mathrm{comp}(\Gamma)$	7.86 / 80.03	< 0.001
$\mathrm{comp}_{K_i}(\Gamma)$	8.05 / 80.03	< 0.001
$\preceq_{lex}$	6.51 / 80.03	< 0.001
$\preceq_{lex_{K_i}}$	2.47 / 80.03	< 0.001

Table 2: Average ranking of the randomly generated statement ϕ_i for each K_i, and p-value for the rankings of all ϕ_i

maximal wrt the lexicographic ordering), such that (1) is the best candidate for removal. So back to the test data, if $K_i = F \cup \{\phi_i\}$, i.e. if ϕ_i is, among the $|F+1|$ statements of K_i, the one which has been randomly generated, and if $\Gamma_{i,i} = K_i \setminus \phi_i$ is among the best ranked immediate subbases of K_i, this would indicate that the linguistic compliance score in definitions 4.3 (resp. 4.4), or the corresponding lexicographic ordering $\preceq_{lex}$ (resp. $\preceq_{lex_{K_i}}$) is actually a good estimator of faulty statements.

An additional precaution was taken in order to avoid artificially good results. For most statements $\phi_j \in K_i$, discarding ϕ_j did not have any impact on the set $\Psi_{\Gamma_{i,j}}$ of consequences to be evaluated, i.e. $\Psi_{\Gamma_{i,j}} = \Psi_{K_i}$, and therefore $\mathrm{comp}(\Gamma_{i,j}) = \mathrm{comp}(K_i)$. Let $\Delta_i \subseteq K_i$ be the set of statements whose removal did have an impact instead (on average, there were 79.3 statements in Δ_i). Then the compliance of a subbase $\Gamma_{i,j}$ of K_i was evaluated only if $\phi_j \in \Delta_i$, i.e. only if the removal of ϕ_j made a difference. K_i was also added to this set of evaluated subbases, yielding a ranking of $79.03 + 1 = 80.03$ bases on average.

Results are again positive. Column "rank" in table 2 gives the average ranking of $\Gamma_{i,i}$, i.e. the base obtained after the removal of the randomly generated statement ϕ_i. Both lexicographic orderings outperformed the compliance scores (i.e. the mean of plausibility scores), and the best configuration was the fourth presented in section 4.2, using $\mathrm{sc}_{K_i}(\psi)$ as a plausibility score instead of $\mathrm{sc}_{\Gamma_{i,j}}(\psi)$.

6 Applications

This section describes a few concrete use cases of the propositions made in section 4. A first basic but useful application is the identification of undesired consequences of a consistent input ontology K. As illustrated by example 1, violations of common sense often go unnoticed in publicly available OWL datasets, even though effective procedures can detect inconsistency[11] in most DLs. This is correlated with the overall sparse usage of negation in OWL, yielding ontologies which are consistent by default rather than by design. The identification of such cases can be very simply performed, by returning to the user the formulas of Ψ_K with lower plausibility scores, like `Person`(*CEO*) in example 1. Axiom pinpointing algorithms (Schlobach and Cornet, 2003; Kalyanpur et al., 2007; Horridge, 2011) may then be used to compute all *justifications* for each returned consequence ψ, i.e. all (set-inclusion) minimal subsets of K which have ψ as a consequence.

In a more automated fashion, the greedy trimming approach described in (Corman et al., 2015) returns n statements of K which are candidate for removal, n being given as a parameter, by incrementally selecting the immediate subbase of Γ with maximal linguistic compliance score, starting with $\Gamma = K$.

But inconsistent[12] ontology debugging may also benefit from distributional evidence. As discussed in section 2, state-of-the-art approaches to ontology debugging suffer from the number of candidate outputs, i.e. of (set-inclusion) maximal consistent subsets of K, as well as from the cost of their computation. If the set $\mathcal{J}$ of justifications for the inconsistency of K is known though, and if some (discriminant enough) preference relation $\preceq_a$ over $\bigcup \mathcal{J}$ can be obtained, then *prioritized base revision*, as it is defined in (Nebel, 1992), provides a principled and computationally attractive solution to these problems. Even if the whole process cannot be depicted here, $\preceq_a$ may actually be obtained through distributional evidence, by evaluating, for each statement $\phi \in \bigcup \mathcal{J}$, the plausibility of some consequences of candidate subbases in which ϕ does or does not appear. The support set in this case is reduced to consequences of the "safe" part of K, i.e. $K \setminus \bigcup \mathcal{J}$.

7 Extensions

A first straightforward extension of this framework consists in taking more complex classes into ac-

[11] or incoherence (see footnote1)

[12] or incoherent (see footnote 1), or for which a set of undesired consequences has already been identified

count. OWL (and most Description Logics) favor the recursive construction of arbitrarily complex classes out of the signature of Γ, and this mechanism could naturally be used to extend Ψ_Γ with more consequences of the form $C(e)$, where C is one of these complex classes. For instance, in example 1, if C_1 and C_2 are respectively defined by $\forall x(C_1(x) \Leftrightarrow \exists y(\texttt{occupation}(y,x))$ and $\forall x(C_2(x) \Leftrightarrow \exists y(\texttt{occupation}(x,y))$, then Ψ_K can be extended "for free" with $C_1(\textit{CEO})$ and $C_2(\textit{Peter Munk})$. Unfortunately, if Ψ_Γ^+ is the set of all consequences of Γ which can be built this way, there is in general no finite subset Ψ_Γ of Ψ_Γ^+ such that $\Psi_\Gamma \models \psi$ for all $\psi \in \Psi_\Gamma^+$. Therefore the complex classes to be used must be selected, which is not trivial. Intuitively, some complex classes are more relevant than other (e.g. the class of "physical objects owned by someone" may be linguistically relevant, but probably not "Moldavian or Muslim lawyers whose father lives in an apartment").

Another simple variation of the framework presented here consists in setting Ψ_Γ to be all consequences of Γ of the form $e_1 \neq e_2$, i.e. the fact that that e_1 and e_2 are not the same individual according to Γ. The unique name assumption is not made in OWL, which means that two distinct named individuals can be interpreted identically, and therefore these consequences do not hold by default. They may be explicitly stated in Γ (owl:differentIndividuals(e_1, e_2)), but are in most cases entailed by Γ, provided it contains some form of negation (e.g. instances of two disjoint classes cannot be the same individual). If Γ_1 and Γ_2 are two subbases of K such that $\Gamma_1 \models e_1 \neq e_2$, but $\Gamma_2 \not\models e_1 \neq e_2$, and if the similarity between e_1 and e_2 is lower than expected, then *ceteris paribus*, Γ_1 will be preferred to Γ_2.

Conclusion

This article is centered on the use of distributional representations of (labels of) named individuals of an input ontology K, in order to identify and repair violations of commonsense within K. For a set of statements $\Gamma \subseteq K$, and Ψ_Γ a specific set of consequences of Γ, a score $\mathrm{sc}_\Gamma(\psi)$ is attributed to each $\psi \in \Psi_\Gamma$, which evaluates the plausibility of ψ wrt Γ according to distributional evidence. Several methods based on this plausibility score are then proposed in order to compare two subbases Γ_1 and Γ_2 of K, leading to the identification of potentially erroneous statements. An evaluation is provided, which consists in extending a test ontology with randomly generated statements before trying to spot them automatically, with significant results. A more thorough evaluation is still required though, testing in particular the impact of a higher number of named individuals and/or classes. Scalability of the approach may also be limited by its heavy reliance on a reasoner. Finally, potential improvements may come from using more elaborated distributional representations, like the one described in (Mikolov et al., 2013).

References

Baroni, M. and S. Bernardini (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC proceedings*.

Benevides, A., G. Guizzardi, B. Braga, and J. Almeida (2010). Validating modal aspects of OntoUML conceptual models using automatically generated visual world structures. *Journal of Universal Computer Science 16*(20).

Buitelaar, P., P. Cimiano, and B. Magnini (2005). *Ontology Learning from Text: Methods, Evaluation And Applications*. IOS Press.

Cimiano, P. (2006). *Ontology learning and population from text: algorithms, evaluation and applications*. Springer.

Cimiano, P. and J. Völker (2005). Towards large-scale, open-domain and ontology-based named entity classification. In *RANLP proceedings*.

Corman, J., N. Aussenac-Gilles, and L. Vieu (2015). Trimming a consistent OWL knowledge base, relying on linguistic evidence. In *LangAndOnto proceedings*.

Ferré, S. and S. Rudolph (2012). Advocatus Diaboli–Exploratory Enrichment of Ontologies with Negative Constraints. *EKAW proceedings*.

Friedrich, G. and K. Shchekotykhin (2005). A general diagnosis method for ontologies. In *ISWC proceedings*.

Giuliano, C. and A. Gliozzo (2008). Instance-based ontology population exploiting named-entity substitution. In *COLING proceedings*.

Horridge, M. (2011). *Justification based explanation in ontologies*. Ph. D. thesis, the University of Manchester.

Kalyanpur, A., B. Parsia, M. Horridge, and E. Sirin (2007). Finding all justifications of OWL DL entailments. In *The Semantic Web*. Springer.

Kalyanpur, A., B. Parsia, E. Sirin, and B. Cuenca-Grau (2006). Repairing unsatisfiable concepts in OWL ontologies. In *ESWC proceedings*.

Mendes, P. N., M. Jakob, and C. Bizer (2012). DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC proceedings*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *ICLR proceedings*.

Nebel, B. (1992). Syntax-based approaches to belief revision. *Belief revision 29*, 52–88.

Pammer, V. (2010). *Automatic Support for Ontology Evaluation*. Ph. D. thesis, Graz University of Technology.

Poveda-Villalón, M., M. C. Suárez-Figueroa, and A. Gómez-Pérez (2012). Did you validate your ontology? OOPS! In *ESWC proceedings*.

Qi, G., P. Haase, Z. Huang, Q. Ji, J. Z. Pan, and J. Völker (2008). A kernel revision operator for terminologies - algorithms and evaluation. In *ISWC proceedings*.

Ribeiro, M. M. and R. Wassermann (2009). Base revision for ontology debugging. *Journal of Logic and Computation 19*(5).

Roussey, C. and O. Zamazal (2013). Antipattern Detection: How to Debug an Ontology without a Reasoner. In *WODOOM 2013 proceeding*.

Schlobach, S. (2005). Diagnosing terminologies. In *AAAI proceedings*.

Schlobach, S. and R. Cornet (2003). Non-standard reasoning services for the debugging of description logic terminologies. In *IJCAI proceedings*.

Suchanek, F. M., M. Sozio, and G. Weikum (2009). SOFIE: a self-organizing framework for information extraction. In *International World Wide Web conference proceedings*.

Tanev, H. and B. Magnini (2008). Weakly supervised approaches for ontology population. In *conference on Ontology Learning and Population proceedings*.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL proceedings*.