Towards Enriching DBpedia from Vertical Enumerative Structures Using a Distant Learning Approach

Mouna Kamel, Cassia Trojahn dos Santos

Towards Enriching DBpedia from Vertical Enumerative Structures Using a Distant Learning Approach

Mouna Kamel and Cassia Trojahn(✉)

Institut de Recherche en Informatique de Toulouse, Toulouse, France
{prenom.nom,cassia.trojahn}@irit.fr

Abstract. Automatic construction of semantic resources at large scale usually relies on general purpose corpora as Wikipedia. This resource, by nature rich in encyclopedic knowledge, exposes part of this knowledge with strongly structured elements (infoboxes, categories, etc.). Several extractors have targeted these structures in order to enrich or to populate semantic resources as DBpedia, YAGO or BabelNet. The remain semi-structured textual structures, such as vertical enumerative structures (those using typographic and dispositional layout) have been however under-exploited. However, frequent in corpora, they are rich sources of specific semantic relations, such as hypernyms. This paper presents a distant learning approach for extracting hypernym relations from vertical enumerative structures of Wikipedia, with the aim of enriching DBpedia. Our relation extraction approach achieves an overall precision of 62%, and 99% of the extracted relations can enrich DBpedia, with respect to a reference corpus.

1 Introduction

In many fields such as artificial intelligence, semantic web or question answering, applications require a reasoning ability, based on semantic resources that describe concepts and relations between. Manually constructing this kind of resource is cost-intensive and results in domain-specific resources of low coverage. However, more than ever automated support for large scale construction of such resources becomes essential. This involves automatically extracting relations from text for building, enriching or populating them. This task usually relies on general purpose corpora as Wikipedia or WordNet and on knowledge extractors mainly exploiting their specific structural elements [22] or sub-corpora [15]. Several of these extractors have targeted these structures in order to enrich or to populate resources as DBpedia [2], YAGO or BabelNet [23].

Enriching DBpedia means identifying new semantic relations from Wikipedia pages. A Wikipedia page is composed of different textual structures which can be divided into three main categories: strongly structured elements, paragraphs which contain plain text, and semi-structured textual units. Strongly structured elements such as infoboxes or User Generated Categories (UGCs) benefit from

https://doi.org/10.1007/978-3-030-03667-6_12

a strong layout, convey a well defined semantics and contain poor written text. Extractors exploiting these elements usually focus on relations (*birthPlace*, *birthDate*, *win-prize*, etc.) which are mostly limited to named entities such as cities, persons, species, etc. With respect to plain text, it has been exploited by numerous relation extraction systems, more often abstracts (whose first sentence is a definition) for identifying hypernym relations[1], other paragraphs for identifying relations in a context of Open Information Extraction. Wikipedia pages are also composed of textual structures, such as titles, subtitles, vertical enumerative structures (i.e. enumerations using typographical and dispositional markers (Fig. 1)). We consider these textual structures as semi-structured ones, as they have the particularity to combine well-written text and layout. Although they express relations which are more often hierarchical relations, these types of structures remain under-exploited as they can not be correctly processed by most classical NLP tools.

The aim of this paper is to show to what extent DBpedia may be enriched with hypernym relations extracted from vertical enumerative structures (VES) present in Wikipedia pages. This kind of relation is central to the construction and enrichment of resources, providing the hierarchical backbone structure of knowledge bases and allows for assigning types to entities. Taking the example in Fig. 1, while several hypernym relations can be identified, e.g., (*Oxfords*,*Men's shoes*) or (*Derby*, *Men's shoes*), few of them are present in BabelNet and none in DBpedia.

We first propose a knowledge extraction approach for identifying hypernym relations carried out by VES. We implement a learning approach for the following reasons (1) the corpus has many regularities that can emerge with this kind of approach and (2) features of different nature (syntactic, lexical, typographical, dispositional, semantic or distributional) can be combined together. In particular, the choice of a distant learning is motivated by the fact that it is free of manual annotation and that the learning knowledge base (here BabelNet) and the learning text (Wikipedia) are aligned, as recommended by the method. We then evaluate the enrichment rate from an experiment we led on a corpus made of VES extracted from French Wikipedia pages.

This work is part of the SemPedia[2] project aiming at enriching DBpedia for French, by specifying and implementing a set of new Wikipedia extractors dedicated to the hypernym relation. We focus on French because semantic resources targeting this language are scarce. We have already proposed a distant supervised approach and implemented a tool for identifying hypernym relations from disambiguation Wikipedia pages [15]. We propose to adapt this approach in this new context, i.e. identifying hypernym relations from vertical enumerative structure of Wikipedia pages, ensuring that the approach is:

– free of manual annotation;

[1] A hypernym relation link two entities E_1 and E_2 when E_2 (hyponym) is subordinate to E_1 (hypernym). From a lexical point of view, this relation is called "isa".

[2] http://www.irit.fr/Sempedia.

Men's shoes can be categorized by how they are closed:

- Oxfords (also referred as "Balmorals"): the vamp has a V-shaped slit to which the laces are attached; also known as "closed lacing". The word "Oxford" is sometimes used by American clothing companies to market shoes that are not Balmorals, such as Blüchers.
- Derby shoe: the laces are tied to two pieces of leather independently attached to the vamp; also known as "open lacing" and is a step down in dressiness. If the laces are not independently attached to the vamp, the shoe is known as a blucher shoe. This name is, in American English, often used about derbys.
- Monk-straps: a buckle and strap instead of lacing.
- Slip-ons: There are no lacings or fastenings. The popular loafers are part of this category, as well as less popular styles, such as elastic-sided shoes.

Fig. 1. Example of VES (https://en.wikipedia.org/wiki/Shoe)

- language independent, thus may be reused for enriching DBpedia in several languages;
- reproducible on any corpus which contains VES having same discourse properties that those of Wikipedia pages.

The rest of this paper is structured as follows. Section 2 presents the background on enumerative structures and on distant learning. We present then our learning model in Sect. 3. Section 4 describes the experimentation and the evaluation. Section 5 discusses the main related work. Finally, Sect. 6 concludes the paper and discusses future work.

2 Background

In this section, we first describe the main principles of distant learning. We then introduce the discursive properties of enumerative structures we lean on to implement our approach.

2.1 Distant Learning

The distant learning method follows the same principles as the supervised learning ones, except that the annotation for constructing the learning examples is carried out using an external semantic resource. In the context of relation extraction from text, it consists in aligning an external knowledge base to a corpus and in using this alignment to learn relations [6,21]. The learning ground is based on the hypothesis that "if two entities participate in a relation, all sentences that mention these two entities express that relation". Although this hypothesis seems too strong, Riedel *et al.* [26] show that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself. Thus, for a pair of entities appearing together within a sentence, a set of features are extracted from the sentence and added to a feature vector for that entity pair.

If the entities are linked in the knowledge base, that entity pair constitutes a positive example, a negative example otherwise.

This approach has been exploited for identifying relations expressed in sentences which are syntactically and semantically correct. Our contribution relies on adapting this approach to different textual structures, especially for those where a part of semantics is carried out by layout. For each type of textual structure, it is then necessary to define a process for building learning examples and to define discriminant features.

2.2 Vertical Enumerative Structures

An enumerative structure (ES) is a textual structure which expresses hierarchical knowledge through different components. According the definition of Ho-Dac *et al.* [12], "it encompasses an *enumerative theme* justifying the union of several elements according to an identity of statut". Different types of enumerative structures exist and different typologies have been proposed.

From a visual point of view, a vertical ES (VES) is expressed using typographic and dispositional markers. More specifically, a VES is composed of (1) a primer (corresponding to a sentence or a phrase) which contains the "enumerative theme" and which introduces (2) a list of items (at least two items) which belong to the same conceptual domain, and (3) possibly of a conclusion. If we consider the example of Fig. 1, "Men's shoes can be categorized by how they are closed:" is the primer, "Oxfords ... such as Blüchers." is an item, *Men's shoes* is the "enumerative theme" and *Oxfords*, *Derby shoes*, *Monk-straps* and *Slip-ons* are entities of the same conceptual domain. This VES has no conclusion.

From a discursive point of view, VES may be classified according to the discourse relations between their components. Before introducing VES properties our approach relies on, we first briefly remind the major principles of the Segmented Discourse Representation Theory (SDRT) [1] which is the discourse theory we used for analyzing VES. A discourse analysis in that context consists in breaking down the text into segments (called discourse units or DU) and in linking adjacent segments with *coordinating* or *subordinating* relations. *Coordinating* relations link entities of the same importance, whereas *subordinating* relations link an entity to an entity of lower importance. Thus, if we consider the primer and items of a VES as DUs (resp. DU_{Primer} and $DU_{Item_j} (j = 1, ..., N)$ if VES is composed of N items), a manual discourse analysis of such a VES allows to state that the primer is linked to the first item with a *subordinating* relation. When all items are linked with *coordinating* relations, we qualify such VES as *paradigmatic* [9] and refer to it as P-VES (Fig. 2(a)).

According again to the SDRT, if DU_{Item_1} is subordinated to DU_{Primer}, hence each DU_{Item_j} coordinated to $DU_{Item_{(j-1)}}$ (j = 2, ..., N), is subordinated to DU_{Primer}. Thereby, N subordinating relations between DU_{Primer} and DU_{Item_j}, (j = 1,...,N), can be inferred (Fig. 2 (b)). In that context and as the *elaboration* relation is a sub-relation of the *subordinating* one, we can also say that each DU_{Item_j} *elaborates* DU_{Primer}. When DU_{Primer} and DU_{Item_j} are broken down into more fine-grained DUs as terms, these discourse relations are kept between

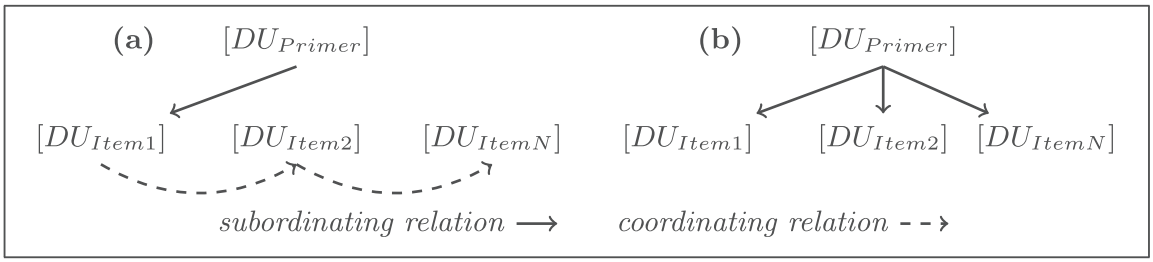

Fig. 2. Discursive representations of P-VES according to the SDRT.

at least one term H in the primer and one term h in the item. From a lexical point of view, these N relations may be specialized in at least N lexical relations $\mathrm{R(H,\ h}_i)_{i=1,\ldots,N}$.

We are interested with P-VES as Wikipedia pages contain many P-VES often expressing definitions and properties of entities. These pages are written according to the guide "The Manual Of Style"[3] which recommends the same grammatical form for all items. An analysis of 100 Wikipedia pages randomly chosen shows that more than 80% of VES respect those instructions and thus are paradigmatic.

We are aware that a P-VES can, however, bear more relations (hierarchical or no hierarchical). The example in Fig. 3 shows that more than one hierarchical relations exist between the primer and the first item which is itself composed of a list (*act of worship* and *sacrifice*, *act of worship* and *libation*, etc.), as well as one no hierarchical (syntagmatic) relation expressed in the last item (*preaching* and *Abrahamic religions*).

The main acts of worship are:

- sacrifice, libation, the offering and education;
- pray (invocation, praise, etc.);
- preaching which plays an important role in Abrahamic religions;

Fig. 3. P-VES containing hierarchical relations and one no hierarchical relation.

3 Proposed Approach

We describe here how the distant learning approach has been adapted for learning hypernym relations from P-VES. We describe in particular the process of building learning examples and the learning model.

[3] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

3.1 Learning Examples Building

The process of building the examples is composed of four main steps, which are detailed in the following.

Step 1: distribution of the primer over the items. On the basis of the discursive properties of P-VES (Sect. 2.2) where each item *elaborates* (in the sense of the *elaboration* discourse relation) the primer, we distribute the primer over each item, giving rise to new textual units called TU. Each *TU* corresponds to the concatenation of the primer and of one item. *N TU* are thus generated if the P-VES is composed of *N* items. For illustrating this step, we consider the P-VES depicted in Fig. 1. Four TU are thus generated:

1. [Men's shoes ... are closed: Oxfords ... such as Blüchers.]
2. [Men's shoes ... are closed: Derby shoe... often used about derbys.]
3. [Men's shoes ... are closed: Monk-straps: ... instead of lacing]
4. [Men's shoes ... are closed: Slip-ons... such as elastic-sided shoes.]

Distributing the primer over the different items does not make us fall back on the classic extraction of relations because we are still confronted with the presence of some typographic markers that replace lexical markers.

Step 2: annotation of terms. In this step, we firstly extract the terms from the external semantic resource (i.e. the terminology provided by the resource, such as synset terms or concept labels). Then, this list is used to annotate the set of generated TUs.

Step 3: building the couples of terms. Several terms may be present in the primer or in an item. Learning examples are then built from couples of terms ($Term_1$,$Term_2$) which respectively belong to the primer and to one of the items. Relying on the *elaboration* relation between the primer and an item, we empirically define the following heuristics for selecting couples of terms:

- $Term_2$ should belong to the first part of the item, i.e. the string starting at index 0 and ending at the next final punctuation (point, line return, etc.) of the item. As Fig. 1, an item may be composed of several sentences.
- Items for which no terms are linked to at least one term of the primer by a hypernym relation, according to the external semantic resource, are left aside. Indeed, this case contradicts our underlying assumption about P-VES that states that each item *elaborates* the primer. This case may nevertheless be explained by a possible incompleteness of the external resource.

For constructing the set of examples, we have generated all the combinations of terms ($Term_1$, $Term_2$) from the primer and retained items. Thus a couple of terms will correspond to a positive example if a hypernym relation between $Term_1$ and $Term_2$ exists in the external semantic resource, to a negative example otherwise. We are aware however that, depending on the coverage of the external semantic resource adopted, negative examples may be false negative ones given the fact that the relation is simply missing in the resource.

Step 4: associating a set of features to couple of terms. Each couple of terms $Term_1$ and $Term_2$ is associated with a set of features coming from the textual units from which they have been extracted. Currently, we focus on lexical features, grammatical features, layout features, and some heuristics inspired by [18], such as the context of $Term_1$ and $Term_2$ (text window). Furthermore, features impact different levels of P-VES: those involving the whole enumerative structure, those involving the example, those involving the primer and those involving the item. Table 1 introduces the set of selected features as used in the experiments described in Sect. 4.

Table 1. Set of learning features (*an enumerative theme is used for organizing the concepts involved into an enumerative structure and is one of the following expressions *list of*, *types of*, *kind of*, etc.).

Scope	Features	Description	Datatype
ES	itemsNumber	number of items present in the VES	integer
Example	lexicalInclusion	lexical inclusion between the terms	boolean
Primer	nbTokens_P	number of tokens in the primer	integer
	lemmaPOSWindow_P	sequence of POS of the window corresponding to 3 tokens preceding $Term_1$, 3 tokens following $Term_1$	string
	lemmaPosTerm1	sequence of POS of all tokens included into $Term_1$	string
	NbTokensBeforeTerm1	number of tokens before $Term_1$	int
	NbTokensAfterTerm1	number of tokens after $Term_1$	int
	capitalizedInitialTerm1	initial of $Term_1$ is capitalized	boolean
	capitalizedTerm1	$Term_1$ is capitalized	boolean
	endsWithColon	primer ends with a colon	boolean
	verbPresence	the primer contains a verbal form	boolean
	theme*	the primer contains an enumerative theme	boolean
	nbTokensTerm1Org	number of tokens between $Term_1$ and the theme	integer
	ordinal	the primer contains a numeral	boolean
	nbTokensTerm1Ord	number of tokens between $Term_1$ and the numeral	integer
Item	nbTokens_I	number of tokens in the item	integer
	nbSentences_I	number of sentences in the item	integer
	lemmaPOSWindow_I	sequence of POS of the window including 3 tokens preceding $Term_2$, 3 tokens following $Term_2$	string
	lemmaPosTerm2	sequence of POS of all tokens included into $Term_2$	string
	NbTokensBeforeTerm2	number of tokens before $Term_2$	int
	NbTokensAfterTerm2	number of tokens after $Term_2$	int
	capitalizedInitialTerm2	initial of $Term_2$ is capitalized	boolean
	capitalizedTerm2	$Term_2$ is capitalized	boolean

3.2 Learning Model

In order to perform a binary classification task (*isA* or *not-isA* classes), we chose the Maximum Entropy classifier (MaxEnt) [3] which is relevant when the conditional independence of the features cannot be assured. This is particularly true in NLP where features are usually words which obviously are not independent in their use (they are bound by syntactic and semantic rules). Furthermore, MaxEnt allows the management of a high number of features. It relies on the maximum entropy principle. Hence, it requires to define a set of constraints for each observation and to choose the distribution which maximizes the entropy

while remaining consistent with the whole set of constraints [14]. In this context of optimisation under constraints, it is mathematically proved that a unique solution exists and that an iterative algorithm converges towards this solution [25]. The classical formula of MaxEnt is the following:

$$P(y|x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(x, y) \right)$$

where $P(y|x)$ gives the probability that the individual x (here a relation) belongs to the class y (here *isA* or *not-isA* classes). Each individual is encoded as a feature vector. The function f_i is a function called *feature* which determines the constraints of the model. The weights w_i associated to each feature account for the probability to belong to a class. Z is a normalization constant which ensures that the sum of probabilities of one individual is equal to 1.

To estimate the parameter values $\hat{w}$, we use the likelihood function that aims at determining the best estimators:

$$\hat{w} = argmax \sum_j log(P(y_j|x_j))$$

where the (x_j, y_j) belongs to the set of training data. We used the OpenNLP (version 1.5.0) implementation of the MaxEnt algorithm[4].

4 Experiments

Our experiments were carried out on the French Wikipedia corpus aiming at enriching French DBpedia, using BabelNet as external semantic resource. These choices are motivated by the fact that (1) semantic resources targeting French language are scarce (French DBpedia is about 20,000 times poorer than DBpedia in English); (2) BabelNet [23] is a multilingual network of concepts and named entities that results from the automatic integration of various background knowledge resources (WordNet, Open Multilingual WordNet, Wikipedia, GeoNames, WoNef, etc.); and the learning knowledge base (BabelNet) and the learning text (Wikipedia) are aligned, as recommended by the method. While in BabelNet some mappings (Wikipedia-WordNet) have been manually checked, YAGO [29] assures a better accuracy of the whole knowledge base. However, we choose BabelNet due to its better coverage of French. It consists of about 14 million entries, including concepts and named entities. Using BabelNet publicly available resource allows for a reproducible approach. We used BabelNet 3.7 version.

4.1 Corpus

We built a corpus from a set of enumerative structures extracted from the 2016 dump of French Wikipedia pages. We have used the WikiExtractor tool[5] for

[4] http://opennlp.apache.org/.
[5] https://github.com/attardi/wikiextractor.

extracting plain text of vertical lists based on HTML tags of lists. We then pre-processed the extracted structures by (1) removing the (multiple) malformed lists; (2) reducing the primer to its last sentence, when the WikiExtractor has extracted a whole paragraph as a primer; (3) reformatting each enumerative structure according to the XML schema we defined; and (5) annotating the corpus with BabelNet label concepts and processing the corpus using Tokenizer, SentenceSplitter, TreeTagger, Gazetteer tools available in the GATE system[6]. This resulted in a corpus of 75446 annotated enumerative structures.

4.2 Learning Examples

From the corpus, 134170 examples were built (2766 positive examples and 131404 negative examples), as the method described in Sect. 3.1:

- an example is positive if $Term_1$ (present in the primer) and $Term_2$ (present in an item) are linked by a direct hypernym relation in BabelNet. We could observe that, given the polysemous nature of terms, considering a high path in the network between them introduce noise. We have then restricted to length 1 the path for classifying the example as positive;
- an example is negative if no path of length lower than 3 exists between $Term_1$ and $Term_2$ in BabelNet. This relaxes the assumption in Sect. 3.1, where examples are assumed to be negative if the relation is simply missing in the resource. From an empirical analysis, we fix to 3 the path length. We could observe that, even if two terms are linked in the resource (with a path length higher than 3), this link may not reflect a hierarchical relation given the polysemous nature of the terms (terms loosely related).

We thus built a training set of 3688 examples (1844 positive and 1844 negative examples) and a test set made of 1844 examples (922 positive and 922 negative examples). Examples have been randomly chosen among the initial sets of positive and negative examples.

4.3 Evaluation Setting

We evaluated our approach on the test set and on 2 reference corpora. These reference corpora have been used by [8] for evaluating their supervised learning approach intended to also identify hypernym relations from enumerative structures. To the best of our knowledge, this is the only corpus of same nature available for comparison. Each reference corpus concerns a set of Wikipedia pages having the same topic, respectively *Computer science* and *Transport*. They have been annotated with terms obtained from both YaTeA and Acabit term extractors. In that work, the results have been reported only in terms of precision on the top 500 hypernym relations extracted. Here, we have used these reference corpora and the reported results as baseline.

[6] https://gate.ac.uk/sale/tao/splitch6.html#chap:annie.

The reference sets correspond to examples built from these annotated reference corpora. We have produced several learning models, varying the different features, and with different sizes of sentence. A linear regression analysis of features gives features listed in Table 1 as discriminant. We keep the model (we named DLM for Distant Learning Model) taking into account these features.

4.4 Results and Discussion

Table 2 presents the results of our approach considering the test set, in terms of precision, recall, F-measure and accuracy. Table 3 presents those considering the reference corpus, in terms of precision. We can observe good values of precision and recall on the test set, while observing varying results in terms of precision on the two reference sets. The low performance of our approach on the *Transport* corpus can be explained by two main reasons. First, this corpus contains several contextual spatial relations, which are expressed using nested VES, where the context is expressed in the primer of one of these nested VES. However, our approach takes the VES independently of each other and hence can not correctly deal with this contextual parameter. Second, these VES are generally composed of numerous items. For the *Computer* corpus our approach outperforms the baseline. Overall, we obtain a precision up to 0.62.

Table 2. Results for the test set for all features of Table 1.

	Precision	Recall	F-measure	Accuracy
DLM on the test set	0.73	0.83	0.78	0.76

Table 3. Results for the reference set for all features of Table 1.

	Precision
DLM on Computer reference set	0.73
DLM on Transport reference set	0.51
Baseline on Computer reference set	0.6
Baseline on Transport reference set	0.6

We could identify some sources of noise in the learning process. First, we could observe that the external resource used here is not exhaustive, which may lead to the generation of false negative examples. That goes against the hypothesis of distant learning approach. Second, false positive examples are introduced due to the fact that the term ambiguity is propagated when exploring the network. Third, the knowledge may be expressed in a different way according to the language. In fact, we can observe cycles in the French network. For instance, the

cycle “microprocessor” is a “microprocessor” in the French network does not exist in the English one.

Besides that, our approach is able to correctly identify the hypernym relations between textual entities (primer, item) that are not contiguous in the text, such as the set of hypernym relations in Fig. 1. These kinds of relations can not be in fact correctly treated by the classical NLP parsers. Furthermore, we can observe that the model is able to correctly identify the cases of head modifiers, and identifies hypernym relations such as (*Ministères*, *Ministère de la Sécurité publique*), (*Ministères*, *Ministère de la Supervision*), (*Ministères*, *Ministère de la Justice*), as from the VES in Fig. 4.

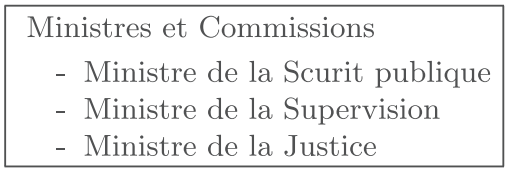

Fig. 4. Hypernym relations identified from head modifiers.

4.5 DBpedia Enrichment

We have evaluated how much our approach could enrich DBpedia with the extracted hypernym relations. These relations contribute to enrich DBpedia in two ways: terms participating in the relations and relations themselves. To do so, we checked the presence/absence in DBpedia of them. The list of checked relations comes from the set of 307 true positive (TP) relations classified by our training model (with a confidence ≥ 0.5) on the reference corpus (168 TP from the *Computer* corpus and 139 TP from the *Transport* corpus).

First of all, we checked how many of the annotated terms are present in DBpedia. For that, we followed two strategies: (i) using a SPARQL query with an exact match between the relation terms and labels of DBpedia resources and (ii) using the DBpedia Spotlight service[7], a tool for automatically annotating mentions of DBpedia resources in texts [7]. We have used the Docker available for French[8]. With the first strategy, for the *Computer* corpus, we found 20 (out of 192) terms in DBpedia, against 9 (out of 168) terms from the *Transport* corpus. Using DBpedia Spotlight, with a confidence of 1 and a support of 20, we found only 3 out of 192 terms from the *Computer* corpus against 6 out of 168 terms from the *Transport* corpus. All of these terms referring to named entities. With a confidence of 0.6, we found 40 terms from the *Computer* corpus and 49 from the *Transport* corpus. However, wrong correspondences have been

[7] https://www.DBpedia-spotlight.org/.
[8] https://github.com/DBpedia-spotlight/spotlight-docker/tree/master/nightly-build/french.

identified lowering the confidence, as somehow expected. This shows that most of the annotated terms from the reference corpus can in fact enrich the resource.

With respect to the 307 TP relations annotated by our model, Table 4 shows the number of relations existing in DBpedia with respect to the presence of their terms in the resource. We can observe that, although some terms participating in the identified relations are present in the resource, only 4 of them participate in the same relation (Spotlight conf = 0.6 in Table 4). Looking at the number of TP relations present in DBpedia, 99% of them are not present in DBpedia. These results confirm that the Wikipedia pages, which are under-exploited by Wikipedia extractors, provide rich hypernym relations other than those found in structured elements (infoboxes, categories, etc.).

Table 4. Presence of relations and their corresponding terms in DBpedia.

	2 terms in DBpedia	Only $Term_1$	Only $Term_2$	None them	Number of present relations
Computer (exact match)	0	17	2	149	0
Transport (exact match)	0	5	3	131	0
Computer (Spotlight conf = 1)	0	6	0	162	0
Transport (Spotlight conf = 1)	0	3	0	136	0
Computer (Spotlight conf = 0.6)	1	35	3	129	0
Transport (Spotlight conf = 0.6)	3	41	2	93	2

5 Related Work

Our approach is related to three main fields of study, whose main related works are discussed in the following.

Enumerative Structures. Firstly, concerning the works on ES, we can mention typologies such as the one proposed by Vergez *et al.* [31], where the items can be present or not in the primer (one-step vs. two-step), or that of Ho-Dac *et al.* [12] where ESs have been classified according to their level of granularity (intra-paragraphic vs. multi-paragraphic). Concerning VES (particularly studied in the context of text generation), Hovy and Arens [13] distinguish the list of items (set of elements of same level) from the enumerated list (for which the order of items is important), while Luc [19] proposes a typology that opposes parallel ES (paradigmatic, visually homogeneous and isolated) to non-parallel ES. This latter is based on the composition of the rhetorical model of Rhetorical Structure Theory (RST) [20] and Textual Architecture Model (TAM) [32]. Drawing inspiration from these works, we also proposed a typology of ES in [9] (we refer to this topology in Sect. 2.2) relying on its discursive properties. These are the discursive properties of paradigmatic VES we exploit in this work.

Hypernym Relation Extraction. Numerous studies have been done in this field and a relevant short overview of them can be found in [33]. The pioneering

work of the linguistic methods is that of Hearst [11] which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted and extended to improve recall for instance with the concept of "star-pattern" [24], or by progressively integrating learning techniques. Snow *et al.* [28] and Bunescu *et al.* [5] apply supervised learning methods to a set of manually annotated examples. While the cost of manual annotation is the main limit of supervised learning, distant learning method consists in building the set of examples using an external resource to automatically annotate the learning examples. For instance, Mintz *et al.* [21] use Freebase as external resource in their distant approach for identifying around 102 different relations. They implement a multi-class optimized logistic classifier using L-BFGS with Gaussian regularization. Another way to avoid manual annotation is the bootstrapping which uses a selection of patterns to construct the set of examples [4]. Some of these works are based on distributional analyses [17].

Enriching Semantic Resources. With respect to the exploitation of Wikipedia for the construction and enrichment of semantic resources, several extractors have been developed to analyze each type of structured data in Wikipedia. Morsey *et al.* [22] developed 19 of such extractors that produce a formal representation of entities and relations identified within various structural elements from Wikipedia: abstracts, images, infoboxes, etc. Other works have targeted specific relations, mainly hypernym relations. For example, Suchanek *et al.* [29] used the User Generated Categories (UGCs) hierarchy of Wikipedia to build hypernym relations in the Yago knowledge base. Kazama and Torisawa [16] exploited the abstract part of the pages, whereas Sumida and Torisawa [30] extracted knowledge from the menu items. Recent works proposed the automatic creation of MultiWiBi [10], an integrated bitaxonomy of Wikipedia pages and categories in multiple languages. Still, extracted relations from the text in Wikipedia pages have been little used to feed DBpedia [27]. Hence, most of the knowledge from these pages remains unexploited.

Here, we target the extraction of hypernym relations from paradigmatic VES (P-VES), which are under-exploited in the literature aiming at enriching existing semantic resources. A previous approach with the same objectives, based on supervised learning and which exploit semantic properties of a P-VES when considering it as a whole semantic unit, have been proposed by [8]. In that context, a precision of about 60% has been obtained. Given the cost-intensive manual annotation of examples, the expertise required for this task, and in order to improve these results, we adopt a different approach based on a distant supervised learning algorithm, and for which P-VES are no longer semantically considered as a whole unit, but are split into N independent textual units (a textual unit is then composed of the primer and one item) if N is the number of items which compose the P-VES. This approach can be carried out on any corpus presenting structural and/or linguistic regularities, such as web documents, and it is language-independent. Indeed, the proposed approach relies on a multilingual resource that can be used for annotating a corpus and on shallow learning features whose extraction does not depend on deep language analyzers.

6 Conclusion and Future Work

This paper has proposed a knowledge extraction approach that exploits vertical enumerative structure, which are frequent in corpora and are rich sources of hypernym relations. They are however under-exploited by knowledge approaches aiming at enriching semantic resources. We applied a distant learning approach for extracting hypernym relations from vertical enumerative structures expressed in French Wikipedia pages, aiming at enriching the French DBpedia. The aim was at evaluating how hypernym relation extraction can take advantage of enumerative structures for enriching existing resources that have been constructed from the same basis. In that sense, we observed that 99% of the extracted relations could be used for enriching the French DBpedia.

As perspectives, we plan to extend our learning features with additional features such as semantic and distributional features and to deal with the disambiguation of terms, as well as to combine different external resources. We intend as well to apply term extractors in order to identify terms before identifying potential relations that can be used for enriching an existing semantic resource. Finally, we plan to exploit different ontology matching methods in order to integrate the extracted relations into the French DBpedia.

References

1. Asher, N.: Reference to Abstract Objects in Discourse: A Philosophical Semantics for Natural Language Metaphysics. SLAP, vol. 50. Kluwer, Dordrecht (1993)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: dbpedia
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Comput. Linguist. **22**(1), 39–71 (1996)
4. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Atzeni, P., Mendelzon, A., Mecca, G. (eds.) WebDB 1998. LNCS, vol. 1590, pp. 172–183. Springer, Heidelberg (1999). https://doi.org/10.1007/10704656_11
5. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724–731 (2005)
6. Bunescu, R.C., Mooney, R.J.: Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, June 2007
7. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
8. Fauconnier, J.P., Kamel, M.: Discovering hypernymy relations using text layout. In: Joint Conference on Lexical and Computational Semantics, Denver, Colorado, pp. 249–258. ACL (2015)
9. Fauconnier, J.-P., Kamel, M., Rothenburger, B.: Une typologie multidimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques. In: Conférence Internationale sur la Terminologie et l'Intelligence Artificielle - TIA 2013, pp. 137–144, Paris, France, October 2013
10. Flati, T., Vannella, D., Pasini, T., Navigli, R.: MultiWiBi: the multilingual Wikipedia bitaxonomy project. Artif. Intell. **241**, 66–102 (2016). (Complete)

11. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, pp. 539–545. Association for Computational Linguistics (1992)
12. Ho-Dac, L.-M., Péry-Woodley, M.-P., Tanguy, L.: Anatomie des Structures Énumératives. In: Traitement Automatique des Langues Naturelles, Montréal, Canada (2010)
13. Hovy, E., Arens, Y.: Readings in intelligent user interfaces. In: Automatic Generation of Formatted Text, pp. 256–262. Morgan Kaufmann Publishers (1998)
14. Jaynes, E.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620 (1957)
15. Kamel, M., Trojahn, C., Ghamnia, A., Aussenac-Gilles, N., Fabre, C.: A distant learning approach for extracting hypernym relations from Wikipedia disambiguation pages. In: International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 6–8 September 2017, France (2017)
16. Kazama, J., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 698–707 (2007)
17. Lenci, A., Benotto, G.: Identifying hypernyms in distributional semantic spaces. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics, pp. 75–79. Association for Computational Linguistics (2012)
18. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: ACL (2016)
19. Luc, C.: Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés. Ph.D. thesis (2000)
20. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
21. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011 (2009)
22. Morsey, M., Lehmann, J., Auer, S., Stadler, C., Hellmann, S.: DBpedia and the live extraction of structured data from Wikipedia. Program Electron. Libr. Inf. Syst. **46**, 27 (2012)
23. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
24. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Stroudsburg, PA, USA, pp. 1318–1327. Association for Computational Linguistics (2010)
25. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania (1998)
26. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
27. Rodriguez-Ferreira, T., Rabadan, A., Hervas, R., Diaz, A.: Improving information extraction from Wikipedia texts using basic English. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)

28. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17 (2004)
29. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge unifying WordNet and Wikipedia. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 697–706 (2007)
30. Sumida, A., Torisawa, K.: Hacking wikipedia for hyponymy relation acquisition. IJCNLP **8**, 883–888 (2008)
31. Vergez-Couret, M., Prevot, L., Bras, M.: Interleaved discourse, the case of two-step enumerative structures. In: Proceedings of Contraints In Discourse III, Postdam, pp. 85–94 (2008)
32. Virbel, J.: Structured Documents, pp. 161–180. Cambridge University Press, New York (1989)
33. Wang, C., He, X., Zhou, A.: A short survey on taxonomy learning from text corpora: issues, resources and recent advances. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1190–1203 (2017)