View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Open Archive Toulouse Archive Ouverte

Open Archive Toulouse Archive Ouverte

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in:
http://oatao.univ-toulouse.fr/22100

To cite this version: Kamel, Mouna and Trojahn, Cassia and Ghamnia, Adel and Aussenac-Gilles, Nathalie and Fabre, Cécile *Extracting hypernym relations from Wikipedia disambiguation pages: comparing symbolic and machine learning approaches.* (2017) In: International Conference on Computational Semantics (IWCS 2017), 19 September 2017 - 22 September 2017 (Montpellier, France).

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr

Extracting hypernym relations from Wikipedia disambiguation pages: comparing symbolic and machine learning approaches

Mouna Kamel[1], Cassia Trojahn[1], Adel Ghamnia[1,2], Nathalie Aussenac-Gilles[1], Cécile Fabre[2]
[1] Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel,cassia.trojahn,adel.ghamnia,nathalie.aussenac-gilles}@irit.fr
[2] Laboratoire CLLE, équipe ERSS, Toulouse, France
cecile.fabre@univ-tlse2.fr

Abstract

Extracting hypernym relations from text is one of the key steps in the construction and enrichment of semantic resources. Several methods have been exploited in a variety of propositions in the literature. However, the strengths of each approach on a same corpus are still poorly identified in order to better take advantage of their complementarity. In this paper, we study how complementary two approaches of different nature are when identifying hypernym relations on a structured corpus containing both well-written text and syntactically poor formulations, together with a rich formatting. A symbolic approach based on lexico-syntactic patterns and a statistical approach using a supervised learning method are applied to a sub-corpus of Wikipedia in French, composed of disambiguation pages. These pages, particularly rich in hypernym relations, contain both kinks of formulations. We compared the results of each approach independently of each other and compared the performance when combining together their individual results. We obtain the best results in the latter case, with an F-measure of 0.75. In addition, 55% of the identified relations, with respect to a reference corpus, are not expressed in the French DBPedia and could be used to enrich this resource.

1 Introduction

In many fields such as artificial intelligence, semantic web, software engineering or information retrieval, applications require a strong reasoning ability, based on semantic resources that describe concepts and the relations between them. These resources can be manually designed. They are of good quality, however due to the high cost of their design, they offer a limited domain coverage. With the increasing amount of textual documents available in digital format, NLP processing chains offer a good support to design such resources from text. In this context, the task of automatically extracting relations from text is a crucial step (Buitelaar et al., 2005). Numerous studies have attempted to extract hypernym relations, as they allow for expressing the backbone structure of such resources and for assigning types to entities.

While symbolic approaches usually rely on manually defined lexico-syntactic patterns identifying clues of relations between terms (Hearst, 1992), statistical approaches, which are nowadays predominant, are generally based on supervised (Pantel and Pennacchiotti, 2008) or unsupervised (Banko et al., 2007) learning methods, or on distributional spaces (Lenci and Benotto, 2012). These methods of different nature answer to the need of exploiting corpora with different specificities (e.g. domain granularity, nature of the corpus, language, target semantic resource, etc.) and which express the hypernym relation in different forms. For giving some examples, this kind of relation can be expressed by the lexicon and the syntactic structure as in the sentence *sand is a sedimentary rock*, by a lexical inclusion as in *domestic pigeon* (implied *domestic pigeon is a pigeon*), or by using punctuation or layout features that replace lexical markers like the comma in *Trojan horse, a Greek myth* or even the disposition in enumerative structures.

The study we conduct in this paper aims to show the interest of applying several approaches on a same corpus in order to identify hypernym relations through their various forms of expression. We are

particularly interested in exploiting a corpus containing both well-written text (i.e., sentences expressed with a complete syntactic structure) and syntactically poor formulations (i.e., sentences with syntactic holes), together with a rich formatting. We analyze the complementarity of a symbolic approach based on lexico-syntactic patterns and a statistical approach based on supervised learning. We applied these two approaches to a corpus of Wikipedia disambiguation pages, which are very rich in hypernym relations differently expressed, as these pages contain both well-written text and poorly-written text.

Our proposal focuses on the combination of the individual results rather than on the combination of the approaches themselves (e.g., by learning patterns). Indeed, combining patterns with machine learning usually relies on path-based methods (Snow et al., 2004) (Snow et al., 2006) (Riedel et al., 2013). However, dependency parsers have proven to perform worse on poorly-written text. Although our approach is naive in that sense, it proves to provide good results, in particular, in terms of F-measure.

This work is part of the SemPedia[1] project that aims at enriching the semantic resource DBPedia for French (semantic resources targeting this language are scarce), by proposing a new Wikipedia extractors dedicated to the hypernym relation. Hence, we evaluate how the extracted relations could potentially enrich such kind of resource.

The paper is organized as follows. Section 2 outlines the main work related to our proposal. Section 3 presents the materials and methods used in our study, namely the description of the training and reference corpus, their pre-processing, and the extraction approaches. The results obtained are presented and discussed in Section 4. Finally, Section 5 concludes the paper and presents future directions.

2 Related work

In the field of relation extraction, the pioneering work of the symbolic methods is that of Hearst (Hearst, 1992) which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted and extended to French for the hypernym relation (Morin and Jacquemin, 2004), for the meronymic relation (Berland and Charniak, 1999), and for different types of relations (Séguéla and Aussenac-Gilles, 1999), by progressively integrating learning techniques.

With respect to statistical approaches and, in particular, those based on machine learning, which are specially required when dealing with large corpus, Snow and colleagues (Snow et al., 2004) and Bunescu and Mooney (Bunescu and Mooney, 2005) apply supervised learning methods to a set of manually annotated examples. While the cost of manual annotation is the main limitation of supervised learning, distant supervision method consists in building the set of examples using an external resource to automatically annotate the learning examples (Mintz et al., 2009). Another way to avoid manual annotation is the semi-supervised learning method called bootstrapping which uses a selection of patterns to construct the set of examples (Brin, 1998). Agichtein and Gravano (Agichtein and Gravano, 2000), and Etzioni and colleagues (Etzioni et al., 2004) have used this method by adding semantic features to identify relations between named entities. Unsupervised learning, based on clustering techniques, was implemented by Yates and colleagues (Yates et al., 2007) and Fader and colleagues (Fader et al., 2011) which used syntactic features to train their classifiers relations between named entities. Some of these works are also based on distributional analyses (Kotlerman et al., 2010) (Lenci and Benotto, 2012) (Fabre et al., 2014). In the work of Kotlerman and colleagues (Kotlerman et al., 2010), they quantify distributional feature inclusion, where the contexts of a hyponym are expected to be largely included in those of its hypernym. Lenci and Benotto (Lenci and Benotto, 2012) explore the possibility of identifying hypernyms using a directional similarity measure that takes into account not only the inclusion of the features of u in v, but also the non-inclusion of the features v in u. The hypothesis from Santus and colleagues (Santus et al., 2014) is that most typical linguistic contexts of a hypernym are less informative than those of its hyponyms.

Beyond these works, which evaluate approaches independently of each other, few results have been reported on the respective contributions and the complementarity of methods. Granada (Granada, 2015) compared the performance of different methods (patterns-based, head-modifier, and distributional ones)

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

for the task of hypernym relation extraction in different languages, by defining several metrics such that density and depth of hierarchies. The evaluation was carried out on different types of corpus but does not take into account the learning approaches. Yap and Baldwin (Yap and Baldwin, 2009) study the impact of the corpus and the size of training sets on the performance of similar supervised methods, on the extraction of several types of relation (hypernym, synonymy and antonymy), whereas Abacha and Zweigenbaum (Ben Abacha and Zweigenbaum, 2011) combine patterns and a statistical learning method based on the SVM classifier for extracting relations between specific entities (disease and treatment) from a biomedical corpus. In the same line, we exploit methods of different nature, focusing on the specific hypernym relation.

In particular, with respect to the approaches combining patterns and distributional methods, most of them rely on path-based methods. It is the case, for instance, of the learning approach of Snow and colleagues (Snow et al., 2004), which automatically learned pattern spaces based on syntactic dependency paths. These paths represent the relationship between hypernym/hyponym word pairs from WordNet and are used as features in a logistic regression classifier. Variations of this method have been applied in different tasks, such as hypernym extraction (Snow et al., 2006) (Riedel et al., 2013) and extraction of definitions (Navigli and Velardi, 2010). However, as stated in (Kotlerman et al., 2010), one major limitation in relying on lexico-syntactic paths is the sparsity of the feature space, since similar paths may somewhat vary at the lexical level. In (Nakashole et al., 2012), generalizing such variations into more abstract paths proved to improve the results, in particular recall. On the other hand, while those approaches mostly rely on dependency trees extracted from well-written text, the performance of dependency parsers has proven to be very low on poorly-written corpus. In that sense, our focus here is to exploit strategies fitting a corpus rich in poorly-written text and where the polysemous occurs frequently (for instance, for the term "Didier Porte", "porte" is tagged as a verb instead of a noun). It is one of the reasons we focus here on the complementarity of the approaches rather than on their combination.

Finally, with respect to the enrichment of DBPedia knowledge base, several tools, called "extractors" have been developed to extract relations from the different elements present in the Wikipedia pages. Morsey and colleagues (Morsey et al., 2012) developed 19 extractors for analyzing abstract, images, infobox, etc. Other works focus on the hypernym relation. For example, Suchanek and colleagues (Suchanek et al., 2007) used the 'Category' part of Wikipedia pages to build the knowledge base, Yago, Kazama and Torisawa (Kazama and Torisawa, 2007) which exploited the 'Definition' part, and finally Sumida and Torisawa (Sumida and Torisawa, 2008) who were interested in the menu items. We can see that the DBPedia knowledge base is built essentially from the structural elements of the Wikipedia pages. Works targeting relation extraction from text have been exploited in a lesser extend (Rodriguez-Ferreira et al., 2016), which means that most of the knowledge in these pages remains under-exploited. Our aim here is to measure the degree of enrichment of semantic resources when exploiting this kind of relation extraction approach.

3 Material and methods

In this section, we describe the Wikipedia sub-corpus we used, its pre-processing, and the extraction methods we have considered.

3.1 Corpus

Different types of pages can be identified within the Wikipedia encyclopedia. Among them, the *disambiguation pages* list the articles whose title is polysemous, giving a definition of all the accepted meanings for this title, which refer to as many entities. Thanks to the Wikipedia's charter guidelines, which recommend the use of templates (for instance, *Toponyms*, *Patronyms*, etc.), these pages present editorial as well as formatting regularities for presenting the different meanings of the term on the page. For each meaning, a definition and a link to the corresponding page are provided. In fact, the definitions are textual objects in which the hypernym relation is often present (Malaisé et al., 2004) (Rebeyrolle

and Tanguy, 2000). Furthermore, on these pages, the definitions take varied but predictable forms. For instance, the following excerpt (Figure 1) which comes from the *Mercure* disambiguation page[2] shows different hypernym relations, expressed thanks to the lexicon (*le mercure est un élément chimique*), with the help of punctuations (the comma in *le Mercure, un fleuve du sud de l'Italie*), taking benefit from the lexical inclusion (*appareil de mesure*, implying that *appareil de mesure* is an *appareil*), or using dispositional and typographical characters (the structure substitutes the lack of complete syntax and expresses a good part of the text meaning) especially when expressing enumerative structure (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.).

Physique et chimie [modifier | modifier le code]

- Le mercure (symbole Hg) est un élément chimique.
- Le terme mercure rouge désignait au XIXe siècle l'iodure de mercure. Dans la dernière partie du XXe siècle il a été appliqué à une substance imaginaire, présentée comme un matériau stratégique rentrant dans la construction des armes nucléaires.
- Le millimètre de mercure (symbole mmHg), ou torr, est unité de mesure de pression.
- Plusieurs appareils de mesure ou méthodes physiques font référence au mercure, dont notamment :
  - la diode à vapeur de mercure,
  - la pile au mercure,
  - la pompe à mercure,
  - le porosimètre à mercure (en),
  - le thermomètre à mercure.

Toponyme et hydronyme [modifier | modifier le code]

Mercure est un nom de lieu notamment porté par :

- Mercure, une station du métro de Lille Métropole ;
- le Mercure, un fleuve du sud de l'Italie ;
- les Îles Mercure, un archipel néo-zélandais, au large de la péninsule de Coromandel.
- le lac Mercure, un lac de l'île principale de l'archipel des Kerguelen, dans les Terres australes et antarctiques françaises ;
- le monastère Saint-Mercure, un important monastère féminin copte orthodoxe, situé dans le vieux Caire (Égypte) ;
- le mont Mercure, une montagne d'Italie ;
- Saint-Michel-Mont-Mercure, une ancienne commune française située dans le département de la Vendée, en région Pays-de-la-Loire
- la Vallée du Mercure, un grand bassin fluvial italien situé dans le sud de la Basilicate et le nord de la Calabre, et qui fut occupé par u lac au Pliocène.

Figure 1: Fragment of the disambiguation page *Mercure*.

We have compiled a corpus made of 5924 French disambiguation pages (XML version of the 2016 Wikipedia dump). From this corpus were extracted two sub-corpora:

- 20 randomly selected disambiguation pages form the *reference corpus*. In these pages, hypernymy relations were manually annotated, marking the terms referring to the related entities and the zone of the text where the relations were identified. This sub-corpus is used to qualitatively evaluate our approach and to evaluate the potential enrichment of DBPedia (Section 4.3);
- the remaining pages form the *training corpus*, which is intended to train and evaluate our learning model (Section 3.3.2).

3.2 Pre-processing

The content of each page has been labeled with morpho-syntactic tags, POS and lemma, using TreeTagger[3]. For identifying the expression of semantic relations, the text is also annotated using terms, namely syntagms, usually nominal, that may designate entities or conceptual classes. For example, *Mercure, système solaire, planète* (*Mercury, solar system, planet*) are some of the terms in Figure 1. The terms can therefore be included in each other (e.g., *system* in *solar system*). Rather than using a term extractor, we chose to construct *a priori* two lists of terms:

[2]https://fr.wikipedia.org/wiki/Mercure

[3]http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger

- LBabel contains the list of terms retrieved from the French labels of concepts present in the semantic resource BabelNet[4]. This list will serve to train the learning system, as detailed in Section 3.3.2;
- LCorpus contains the list composed of the manually annotated terms from the reference corpus (Section 4).

These lists are then respectively projected on the pre-processed learning and reference corpora. In fact, the annotation of the corpus by terms derived from a shared semantic source ensures the validity of the learning model. This also prevents the identification of terms biasing the relation extraction process.

3.3 Relation extraction approaches

As already stated before, we have chosen two approaches of different nature which are often opposed by the cost of their implementation and by the precision and recall they provide: a symbolic approach based on lexical-syntactic patterns, and a statistical approach based on supervised learning using the distant supervision principle. While patterns represent recurring language patterns expressed through lexicon, syntactic and punctuation elements, automatic learning allows for combining features of different natures (morphological, syntactic, semantic or shaping) and for capturing the properties of contexts in a more global way. These approaches are detailed below.

3.3.1 Lexico-syntactic patterns

A lexico-syntactic pattern is a regular expression composed of words, grammatical or semantic categories, and symbols aiming to identify textual segments which match this expression. In the context of relation identification, the pattern characterizes a set of linguistic forms whose the interpretation is relatively stable and which corresponds to a semantic relation between terms (Rebeyrolle and Tanguy, 2000). Patterns are in fact very efficient, particularly in terms of precision, as they are adapted to the corpus. However, since their development is cost-expensive, it is conventional to implement generic patterns such as those of Hearst (Hearst, 1992). Here, we use a more complete list of 30 patterns from the work of Jacques and Aussenac (Jacques and Aussenac-Gilles, 2006)[5]. We have also extended this set of patterns with more specific (ad-hoc) patterns which better fit the template structure of disambiguation pages (Ghamnia, 2016). This set of enriched patterns are the one used in our experiments.

3.3.2 Distant supervision learning

We have chosen to use the principle of distant supervision proposed by Mintz and colleagues (Mintz et al., 2009). This approach consists in aligning an external knowledge base to a corpus and in using this alignment to learn relations. 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 and colleagues (Riedel et al., 2010) have showed that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself.

As with any supervised learning method, it is necessary to create a set of examples, to train a statistical model on these examples, and to evaluate the model on a test set or by cross-validation. The originality of this approach refers to the fact that the learning examples are automatically built with the help of a semantic resource: the class associated to a pair of terms present in a same sentence, corresponds to the the relation (if it exists) that binds these terms in the external resource. Once trained from the learning examples, a multi-class classification algorithm makes it possible to associate a class (and therefore a relation) with each example of a new corpus.

We have adapted this method by focusing on the hypernym relation, and proceeding to a binary classification. A pair of terms is classified as a positive (negative) example if the two terms denoting two concepts that exist in the semantic resource are linked (are not linked) with the hypernymy relation

[4]http://babelnet.org/

[5]A JAPE implementation of these two types of patterns is available on https://github.com/aghamnia/SemPediaPatterns

in this resource. In all other cases, the term pair is not an example of learning. Our learning examples are constructed with reference to the semantic resource BabelNet which has the advantage of integrating various knowledge bases including DBPedia, the semantic resource that we want to enrich in a long term. In addition, the hypernymy relation is more straightforward expressed in BabelNet than in DBPedia.

Each example is built from a context which encompasses the two terms that are possibly linked by a relation. A context (or window) consists in n (n being the size of the window) tokens preceding, following and separating the two terms. The features are then extracted from that context. These features are described in Table 1.

Scope	Features	Signification	Type
Token	POS	Part Of Speech	string
	lemma	Lemmatized form of the token	string
Window	distT1	Number of tokens between the token and Term1	integer
	distT2	Number of tokens between the token and Term2	integer
	nbMotsFenłtre	Number of tokens in the window	integer
	distT1T2	Number of tokens between Term1 and Term2	integer
Sentence	nbMotsPhrase	Number of tokens in the sentence	integer
	presVerbe	Presence of a verbal form	boolean

Table 1: Features set.

Although the features we use here do not take into account more sophisticated structures, as dependency trees (as discussed in Section 2), they provide quite good results, as discussed in the next sections.

We illustrate the content of a feature vector with the following example where the length of the window is fixed to 3 (we have evaluated windows of dimensions 1, 3 and 5, the optimum being obtained for length 3):

"Lime ou citron vert, le fruit des limettiers : Citrus aurantiifolia et Citrus latifolia"

Mapping the list of terms leads to annotate the sentence with terms `Lime`, `citron`, `citron vert`, `vert`, `fruit`. Let us consider the pair <`Lime`, `fruit`> randomly chosen by the system: Term1=`Lime` and Term2=`fruit`.

The system thus extracts:

```
Terme1 ou citron vert, le Terme2 des limettiers :
```

where tokens corresponding to terms have been replaced with Term1 and Term2. TreeTagger annotation allows to replace the exact form of tokens by their part of speach followed by their lemma:

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
```

Finally, feature functions give distances (in number of tokens) between a token and the annotated terms in the form of the pair of values, the number of tokens between Term1 and Term2 (here 5) and the number of tokens in the whole sentence (here 16). The last feature indicates the presence of a verbal form to discriminate poorly-written text from well-written text.

```
(1,-5) (2,-4) (3,-3) (4,-2) (5,-1) (7,1) (8,2) (9,3) 5 16 true
```

The entire example leads to the following representation:

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2
PRP:det/du NOM/limettier PUN/:
(1,-5) (2,-4) (3,-3) (4,-2) (5,-1) (7,1) (8,2) (9,3) 5 16 true
```

This example is a positive one as a hypernym link between *lime* and *fruit* exists in BabelNet.

From the whole set of examples produced according to the process described above, we randomly selected 3000 positive examples and 3000 negative examples (from a total of 84169 examples). From these 6000 examples, 4000 are used as the set of training examples and 2000 form for the test set (with a rate of 50% of positive examples, for both training and test sets). We are aware that the strategy we follow to split the set of examples may affect the results. Although alternative strategies consider, for instance, the zero-lexical overlapping, as adopted by Weeds and colleagues (Weeds et al., 2014) and Levy and colleagues (Levy et al., 2015), we can not follow this kind of strategy here due to the nature of the corpus, where each sentence of a page corresponds to a characterization or a definition of the entity described by this page.

We have trained a binary logistic regression algorithm, the Maximum Entropy classifier MaxEnt (Berger et al., 1996) on the training set. When applying this algorithm on the test set, we obtained a recall of 0.63 and an accuracy of 0.71.

4 Results and discussion

In the following, we discus the results of the approaches described above and we evaluate their complementarity, as well as the advantage of combining their results.

4.1 Results

The quantitative evaluation we present in this section is not intended to measure the performance of the approaches in absolute terms, but rather to know the order of magnitude of the number of relations found by each of them, whether they are common or specific. This evaluation is based on the reference corpus. The set of examples from the reference corpus contains 688 true positive examples (TP) and 267 true negative examples (TN). We consider the relations extracted by each of the approaches as well as the intersection and the union of the set of relations. Table 2 provides the results in terms of precision, recall, F-measure and accuracy. We can observe that we obtain the best values of F-measure when combining both the results of patterns and MaxEnt.

	Patterns	MaxEnt	Patterns inter MaxEnt	Patterns union MaxEnt
Precision	0.81	0.71	0.75	0.73
Recall	0.46	0.63	0.32	0.77
F-measure	0.53	0.67	0.45	0.75
Accuracy	0.54	0.55	0.43	0.63

Table 2: Evaluation of the approaches.

As we will better discuss in the next section, the two approaches do not often find the same relations, what corroborates their complementarity.

4.2 Discussion

We have carried out an analysis of the nature of the differences in the set of extracted relations from both approaches. We first counted the number of relations found by each approach individually, by both of them, or by none of the two (Table 3), with respect to the true positive relations extracted from the reference corpus.

Among the 221 TP relations found by the two approaches, few of them (9 relations) are expressed by the verb *to be*, as for instance, between the terms *macédoine* (*macedonia*) and *salade de fruits* (*fruit salad*) in the sentence “La macédoine est une salade de fruits ou de légumes” (*Macedonia is a salad of fruit or of vegetables*). Almost all other relations correspond to the pattern “X, Y” as in the sentence “Le cheval de Troie, un mythe grec” (*The Trojan horse, a Greek myth*).

	Number TP
Found by patterns AND MaxEnt	221
Found by patterns AND NOT by MaxEnt	96
Found by MaxEnt AND NOT by patterns	210
Found neither by MaxEnt nor by patterns	161

Table 3: Number of true positives (TP) found by the approaches.

From the 96 relations found by patterns and which were not identified by MaxEnt, 19 of them are expressed with the help of the verb *to be*, in particular when the relation is not expressed at the beginning of the sentence, as the relation between *Babel fish* and *espèce imaginaire* (*imaginary species*) in the sentence "Le poisson Babel ou Babel fish est une espèce imaginaire" (*Babel fish or Babel fish is an imaginary species*). Most of the remaining relations match again the pattern "X, Y". We observe as well that the cause of the silence of MaxEnt in this case may be some specific syntactic variations, such as the presence of dates between parentheses, different punctuation, etc. Indeed, our learning model is sensitive to these variations as sentences are very short and present strong regularities.

Among the 210 relations found by MaxEnt, and not found by patterns, we can observe that (i) many relations are expressed with the help of a lexical inclusion, as for instance in the noun phrase *gare de Paris Bastille* (*Paris Bastille railway station*) used to identify the relation *gare de Paris Bastille* (*Paris Bastille railway station*) is a *gare* (*railway station*); (ii) some relations are expressed with the help of a state verb, as for the relation between *aigle* (*eagle*) and *oiseaux* (*birds*) in the sentence "Aigle désigne en fran{cais certains grands oiseaux rapaces" (*Eagle designates some large birds*). We can notice as well that MaxEnt is able to identify the relations expressed in textual units containing a coordination, as the relation between *poisson Babel* (*Babel fish*) and *espèce imaginaire* (*imaginary specie*) in the sentence "Le poisson Babel ou Babel fish est une espèce imaginaire", or between *Louis Babel* and *explorateur* (*explorer*) in the sentence "Louis Babel, prêtre-missionnaire oblat et explorateur" (*Louis Babel, oblate missionary priest and explorer*). Finally, MaxEnt is also able to identify the relations within the text using formatting as in the relation between *arête* (*ridge*) and *barbe de l'épi* (*beard of the ear*) in the sentence "Arête, "barbe de l'épi"" (*Ridge, 'beard of the ear'*) or between *Aigle* and *chasseur de mines* (*mine hunter*) in the sentence "Aigle (M647), chasseur de mines" (*Eagle (M647), mine hunter*).

From the 161 true positive relations missed by both patterns and MaxEnt, 64 are expressed in sentences that contain parenthetical clauses which separate two terms, as in the sentence "Un Appelant (jansénisme) est, au XVIIIe siècle, un ecclésiastique qui appelle ... (*An Appellant (jansenism) is, during the XVIIIth century, an ecclesiastic who calls ...*) where the relation *Appelant* (*Appellant*) is a *ecclésiastique* (*ecclesiastic*) is not found. 55 of them correspond to the relations expressed by head modifier. We could not precisely identify the silence of MaxEnt in this case. The remaining 42 cases concern forms of expression not supported by the patterns and too scarce to be learned by MaxEnt, such as "X such as Y".

Furthermore, we could also observe that patterns were able to identify relations between common names, rather than between named entities, whereas MaxEnt mainly finds relations between named entities. The reason is that some patterns identify phrases that may not be annotated with the LCorpus terms.

In summary, these results corroborate the gain brought by the combination of complementary methods on the same corpus. Firstly, we could notice that MaxtEnt is able to identify hypernym relations within complex phrases or textual structures, such as vertical item lists, provided they appear with a minimal frequency. Secondly, the different occurrences of relations within the same sentence are identified by the two methods, as seen above through the example "Le poisson Babel ou Babel fish est une espèce imaginaire". In these experiments, patterns and MaxEnt are complementary in a proportion of ˜1/3 vs. 2/3.

4.3 DBPedia enrichment

In a last stage, we evaluated how much our approach could enrich DBPedia with the extracted relations. To do so, we manually checked their presence/absence in DBPedia. This verification had to be manual because the annotated terms come from LCorpus and may differ from the labels in DBPedia. We queried DBPedia to check if entities with labels close to *Term1* and *Term2* were linked by a path made of *rdf:type* and *rdf:subclassOf* relations. We set to 3 the maximum path length.

From the 688 TP in the reference corpus, 199 relations were not expressed in DBPedia. 103 of these 199 relations were identified by the learning approach and 42 of them were found by patterns. Considering the union of the results of the two approaches, 125 identified relations were not in DBPedia (20 relations belonging to the intersection of the individual results). Table 4 presents the rate of enrichment of DBPedia with respect to the relations identified by each approach and the union of their results. These figures confirm that the Wikipedia text, which is under-exploited by Wikipedia extractors, contains hypernym relations other than those found in structured elements (infobox, categories, etc.).

Method	Enrichment rate
Patterns	21%
MaxEnt	51%
Pattern union MaxEnt	63%

Table 4: DBPedia enrichment rate.

5 Conclusion and perspectives

The study reported in this paper led us to set up a methodology to compare two relation extraction approaches, in order to analyze their complementarity. The first results are encouraging and converge with the work of (Malaisé et al., 2004) (Granada, 2015) (Buitelaar et al., 2005). We plan to push this research further on in several directions. We want to integrate other methods, taking into account other textual elements, for example the system of (Kamel and Trojahn, 2016) that deals with vertical and regular enumerative structures, or the tools developed in (Granada, 2015). For improving the performance of each method, in addition to a better pattern encoding, we plan to add new features to the learning process. Moreover, the method will have to be tested on another corpus including other types of Wikipedia pages.

Ultimately, our ambition is to cross the methods so that the results of some serve as richer inputs to others, and thus improve their performance. The first step in this direction would be to annotate the corpus using patterns and tag it to signal whether a pattern is (or is not) recognized in the context of two terms, which would be a strong sign of the presence of the relation. This type of feature would allow the classifier to recognize several types of relations in addition to hypernymy.

Acknowledgement

The authors would like the Midi-Pyrénées (now Occitanie Pyrénées-Méditerranée) Region who funded the SemPedia project and Adel Ghamnia's Ph.D. grant.

References

Agichtein, E. and L. Gravano (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM conference on Digital libraries*, pp. 85–94. ACM.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open information extraction from the web. In *IJCAI*, Volume 7, pp. 2670–2676.

Ben Abacha, A. and P. Zweigenbaum (2011). A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts. In A. Gelbukh (Ed.), *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing, Tokyo, Japan, February 20-26*, pp. 139–150. Springer Berlin Heidelberg.

Berger, A. L., V. J. Della Pietra, and S. A. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics 22*(1), 39–71.

Berland, M. and E. Charniak (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 57–64. Association for Computational Linguistics.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pp. 172–183. Springer.

Buitelaar, P., P. Cimiano, and B. Magnini (2005). Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, pp. 3–12. IOS Press.

Bunescu, R. C. and R. J. Mooney (2005). 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. Association for Computational Linguistics.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2004). Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pp. 100–110. ACM.

Fabre, C., N. Hathout, L.-M. Ho-Dac, F. Morlane-Hondère, P. Muller, F. Sajous, L. Tanguy, and T. Van De Cruys (2014, June). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France, pp. 196–205.

Fader, A., S. Soderland, and O. Etzioni (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545. Association for Computational Linguistics.

Ghamnia, A. (2016). Extraction de relations dhyperonymie partir de wikipdia. In *Actes de la confrence conjointe JEP-TALN-RECITAL 2016*.

Granada, R. L. (2015). *Evaluation of methods for taxonomic relation extraction from text*. Ph. D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, Stroudsburg, PA, USA, pp. 539–545. Association for Computational Linguistics.

Jacques, M.-P. and N. Aussenac-Gilles (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thmatique 47*(1).

Kamel, M. and C. Trojahn (2016). Exploiter la structure discursive du texte pour valider les relations candidates d'hyperonymie issues de structures énumératives parallèles. In *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances, Montpellier, France, June 6-10, 2016.*, pp. 111–122.

Kazama, J. and K. Torisawa (2007). 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.

Kotlerman, L., I. Dagan, I. Szpektor, and M. Zhitomirsky-geffet (2010, October). Directional distributional similarity for lexical inference. *Nat. Lang. Eng. 16*(4).

Lenci, A. and G. Benotto (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 75–79. Association for Computational Linguistics.

Levy, O., S. Remus, C. Biemann, and I. Dagan (2015). Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*.

Malaisé, V., P. Zweigenbaum, and B. Bachimont (2004). Detecting semantic relations between terms in definitions. In *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pp. 55–62. COLING.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). 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: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics.

Morin, E. and C. Jacquemin (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities 38*(4), 363–396.

Morsey, M., J. Lehmann, S. Auer, C. Stadler, and S. Hellmann (2012). Dbpedia and the live extraction of structured data from wikipedia. *Program 46*(2), 157–181.

Nakashole, N., G. Weikum, and F. Suchanek (2012). Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 1135–1145. Association for Computational Linguistics.

Navigli, R. and P. Velardi (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 1318–1327. Association for Computational Linguistics.

Pantel, P. and M. Pennacchiotti (2008). Automatically harvesting and ontologizing semantic relations. *Ontology learning and population: Bridging the gap between text and knowledge*, 171–198.

Rebeyrolle, J. and L. Tanguy (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire 25*, 153–174.

Riedel, S., L. Yao, and A. McCallum (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163.

Riedel, S., L. Yao, A. Mccallum, and B. M Marlin (2013). Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL-HLT 2013*.

Rodriguez-Ferreira, T., A. Rabadan, R. Hervas, and A. Diaz (2016, may). Improving Information Extraction from Wikipedia Texts using Basic English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Santus, E., A. Lenci, Q. Lu, and S. Schulte im Walde (2014). Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pp. 38–42. Association for Computational Linguistics.

Séguéla, P. and N. Aussenac-Gilles (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Conférence ingénierie des connaissances*, pp. 79–88.

Snow, R., D. Jurafsky, and A. Y. Ng (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Snow, R., D. Jurafsky, and A. Y. Ng (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 801–808.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pp. 697–706.

Sumida, A. and K. Torisawa (2008). Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, Volume 8, pp. 883–888.

Weeds, J., D. Clarke, J. Reffin, D. J. Weir, and B. Keller (2014). Learning to distinguish hypernyms and co-hyponyms. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 2249–2259.

Yap, W. and T. Baldwin (2009). Experiments on pattern-based relation learning. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1657–1660. ACM.

Yates, A., M. Cafarella, M. Banko, O. Etzioni, M. Broadhead, and S. Soderland (2007). Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 25–26. Association for Computational Linguistics.