International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

A Distant Learning Approach for Extracting Hypernym Relations from Wikipedia Disambiguation Pages

Mouna Kamel[a,b,*], Cassia Trojahn[b], Adel Ghamnia[b,c], Nathalie Aussenac-Gilles[b], Cécile Fabre[c]

[a] *Université de Perpigan, France*
[b] *IRIT, CNRS, Université de Toulouse, France*
[c] *CLLE, équipe ERSS, Université de Toulouse, France*

Abstract

Extracting hypernym relations from text is one of the key steps in the automated construction and enrichment of semantic resources. The state of the art offers a large varierty of methods (linguistic, statistical, learning based, hybrid). This variety could be an answer to the need to process each corpus or text fragment according to its specificities (e.g. domain granularity, nature, language, or target semantic resource). Moreover, hypernym relation may take different linguistic forms. The aim of this paper is to study the behaviour of a supervised learning approach to extract hypernym relations whatever the way they are expressed, and to evaluate its ability to capture regularities from the corpus, without human intervention. We apply a distant supervised learning algorithm on a sub-set of Wikipedia in French made of disambiguation pages where we manually annotated hypernym relations. The learned model obtained a F-measure of 0.67, outperforming lexico-syntactic pattern matching used as baseline.

Keywords: Distant learning, hypernym relation, knowledge extraction, Wikipedia, knowledge bases

1. Introduction

Extracting hypernym relations from text is one of the key steps in the automated construction and enrichment of semantic resources, since this kind of relation provides the hierarchical backbone structure which allows for entity type assignment. Several hypernym extraction methods have been proposed in the literature, trying to better identify the different ways this kind of relation is expressed in written natural language.

While linguistic methods rely on lexico-syntactic patterns used to identify clues of relation between terms [1], statistical methods, which are predominant, rely on distributional spaces [2] or on supervised [3] or non-supervised [4] learning methods. Most of these technique apply on well-written text (i.e., where the language syntactic structure is well-formed), without taking into account its structure. However, some of them exploit specific textual structures, e.g. definitions [5], enumerative structures [6] or elements like infoboxes, categories or links [7] in the case of Wikipedia.

* Corresponding author.
E-mail address: mouna.kamel@irit.fr

DOI: 10.1016/j.procs.2017.08.208

These methods have been developed for corpora with their own specificities, e.g., domain granularity (general or specific), corpus gender (encyclopedic, scientific, journalistic, etc.), language, explicitness of the text structure (structured, semi-structured or unstructured). Another design parameter is the intended aim (i.e. extracting linguistic relations or formal triples, annotating text or populating a knowledge base). When relations have to be included in a semantic resource, the nature of this resource (thesauri, termino-ontologies or heavy ontologies) also impacts the extraction process. In addition, whatever the input corpus, hypernym relations may be expressed in different forms.

In this paper we study the behaviour of a distant supervised learning approach[8] on a corpus where the hypernym relation is expressed in different forms. Supervised learning algorithms carry out classification based on features of the entities to be classified. When applied to text, these features include various linguistic clues (either syntactic, semantic, lexical, visual, structural, distributional clues). Training the algorithm requires the corpus to be annotated with examples of class to be learned, which is complex and time consuming. Distant supervision overcomes this limitation by relying on an external semantic resource that is mapped to the corpus to automatically generate relation annotations. Being free of manual annotation, this approach can easily be applied to any corpus with regular structures.

This work is carried out within the SemPedia[1] project whose goal is to enrich DBPedia[2] 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. The French DBPedia resource is 20,000 times poorer than DBPedia in English. Then, we built a corpus made of Wikipedia disambiguation pages in French. These pages list the Wikipedia articles whose title is ambiguous, and give a definition of each accepted meaning. Therefore, they are rich in named entities and in hypernym relations expressed through textual definitions and entity types. Moreover, we observed that pages are structured at various degrees depending on the language. For example, pages in English are very highly structured and mostly contain "low-written text" (e.g., opposed to well-written text), where the structure substitutes the lack of full syntax and expresses a good part of the text meaning. The French pages, on the contrary, more readily mix written text and low-written text. So they are a favorable case of relation-rich pages.

Concerning the automatic annotation, we used BabelNet[9], a knowledge resource which is (partially) derived from the training Wikipedia corpus, as recommended by the distant supervision approach. We compare our approach to a symbolic one based on lexico-syntactic patterns, which have been largely used to identify hypernym relations, in particular, on corpora rich in definitions[1], such as the corpus we used in our experiments.

This paper is structured as follows. Section 2 deals with the main related work. Section 3 introduces the distant supervised learning approach and the Maximum Entropy algorithm that we use. Then we present the hypernym relation learning task in Section 4. Section 5 describes the experiments and discusses the obtained results. Finally, Section 6 concludes the paper and draws perspectives for future work.

2. Related work

In the field of relation extraction, the pioneering work of the linguistic methods is that of Hearst[1] which defined a set of lexico-syntactic patterns specific to the hypernym relation for English. This work has been adapted to French in order to identify different types of relations[10], hypernym relations between terms[11] or meronymic relations[12]. Moreover paterns have been progressively learned from text thanks to learning techniques.

With respect to statistical approaches, Snow and colleagues[13] and Bunescu and colleagues[14] applied supervised learning techniques on a set of manually annotated examples. Because the cost of manual annotation is one of the main limitations of supervised learning, distant supervision learning consists in building the set of examples thanks to an external resource[8]. Distant supervision avoids the manual annotation phase by matching relations from the external knowledge base on the corpus. This process allows to automatically annotate relation occurrences that will become learning examples. Another way to avoid manual annotation has been proposed by Brin[15] who uses a selection of patterns to construct the set of examples, thanks to a semi-supervised learning method called bootstrapping. Agichtein and Gravano[16], and Etzioni and colleagues[17] have used this method and added semantic features to identify relations between named entities. IAn alternative technique is unsupervised learning based on clustering techniques. Yates and

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

[2] DBPedia is a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Linked Open Data

colleagues[18] and Fader and colleagues[19] implemented unsupervised learning and used syntactic features to train their classifiers on relations between named entities. Some of these works are also based on distributional analyses[2,20]. To better understand how to take into account the specificities of the corpora, Yap and Baldwin[21] studied the impact of the corpus and the size of training sets on the performance of supervised methods for the extraction of different types of relation (hypernym, synonymy and antonymy). Granada[22] compared the performance of different methods (patterns-based, head-modifier, and distributional ones) for the task of hypernym relation extraction on various kinds of corpora (encyclopedic, journalistic) in several languages.

Relation extraction can also take advantage of the page layout in two different ways. The first one relies on documents written in a markup language. The semantics of the tags and their nested structure maybe exploited for the identification of relations. A collection of XML documents has been used to build ontologies[23,24], while a collection of HTML or MediaWiki documents has been processed to build taxonomies[25]. The second category of approaches bears on specific documents or parts of them, for which the layout defines a precise semantics, such as dictionaries and thesaurus[26] or specific and well localized textual structures such as tables, categories[27,28] or infoboxes[29] from Wikipedia pages. Any of these textual structures can also be made explicit thanks to a markup language.

Finally, as our study focuses on the Wikipedia corpus, the extracted relations could be exploited to enrich DBPedia. Several tools, called "extractors" have been developed to analyze each type of structured data in Wikipedia. Morsey and colleagues[7] developed 19 of such extractors that produce a formal representation of entities and relations identified within various structural elements from Wikipedia: abstracts, images, infobox, etc. Other works have targeted specific relations, mainly hypernym relations. For example, Suchanek and colleagues[28] used the 'Category' hierarchy of Wikipedia to build hypernym relations in the Yago knowledge base. Kazama and Torisawa[30] exploited the 'Definition' part of the pages, whereas Sumida and Torisawa[25] extracted knowledge from the menu items. Recent works proposed the automatic creation of MultiWiBi[31], an integrated bitaxonomy of Wikipedia pages and categories in multiple languages. Still, relation extraction from the text in Wikipedia pages has been little used to feed DBPedia[32]. Hence, most of the knowledge from these pages remains unexploited. This lack is even more important for pages in French.

This context led us to define methods that could extract knowledge from the text in French Wikipedia pages. As a first step towards this goal, we target the extraction of hypernym relations whatever the way they are expressed, be it in well-written or in low-written text. Given the size of the corpus and the cost and expertise by the manual annotation of examples, we apply a distant supervised learning algorithm. We also decided to combine various features, at different linguistic levels (from morphology and syntax to discourse and layout). This approach can be carried out on any corpus presenting structural and/or linguistic regularities, such as web documents, and it is language-independent.

3. Background

3.1. Distant supervision learning

Distant supervision learning[33],[8] refers to learning algorithms where the training examples are automatically collected using a knowledge base. The set of examples is built by aligning the knowledge base to a corpus. The resulting alignments (or text annotations) and their features are then used to train the system and 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[34] show that it makes sense when the knowledge base used to annotate the corpus is derived from the corpus itself. Mintz and colleagues[8] use Freebase as external resource. For every pair of entities linked in Freebase and appearing together within a sentence, a positive learning example is built, i.e., the learning features are extracted from the sentence and added to a feature vector for that entity pair. The set of feature vectors feed a multi-class logistic regression classifier. While Mintz and colleagues[8] consider several relations at once with a multi-class classifier, given the size of our corpus, we only focus on the hypernym relation and a binary logistic regression classifier, the Maximum Entropy Algorithm.

3.2. The Maximum Entropy algorithm

To perform the binary classification task (*isA* or *not-isA* classes), we chose the Maximum Entropy classifier (MaxEnt)[35] which is relevant when the conditional independence of the features can not 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 sematic rules). Furthermore, MaxEnt allows the management of a great 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[36]. 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[37].

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. In our work, we used the OpenNLP (version 1.5.0) implementation of the MaxEnt algorithm[3].

4. Hypernym relation learning task

In this section, we firstly describe the resources we used for training the models. Then, these models are presented.

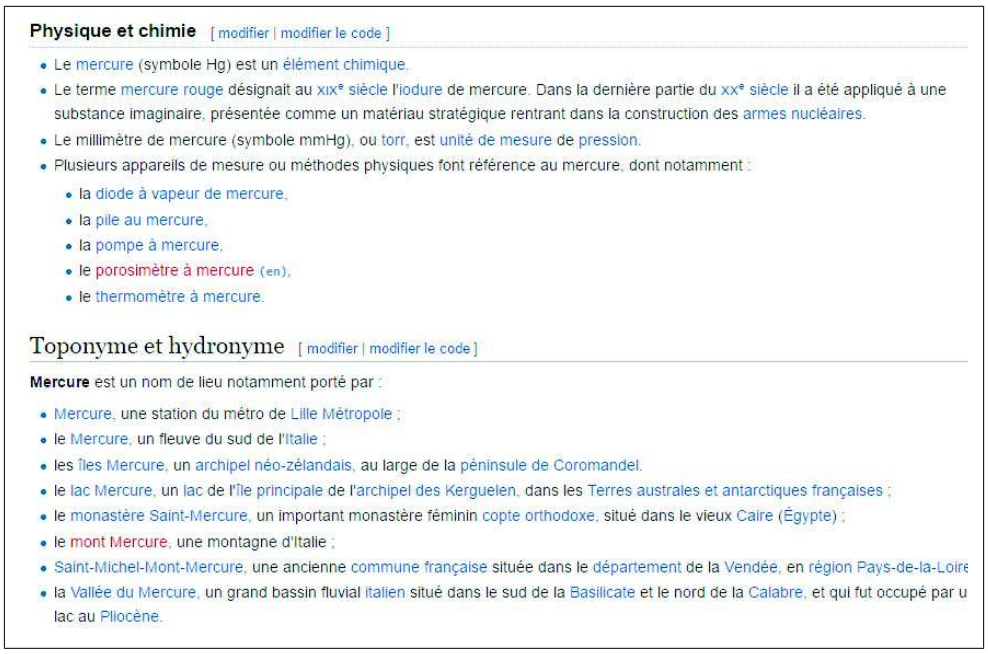
4.1. Resources

We use the Wikipedia sub-corpus composed of all French disambiguation pages. These pages list the Wikipedia articles whose title is ambiguous, and give a definition of each accepted meaning. Therefore they are rich in named entities and in hypernym relations. Relations are expressed in different textual structures, usually established by the guidelines of the drafting charter. These HTML pages are semi-structured textual documents that combine different levels of text structuring, translated by typographical and layout features. The combination of these feature lead to a large variety of possibilities to express hypernym relations. Figure 1 presents the *Mercure* disambiguation page[4], where hypernym relations are expressed thanks to the following linguistic elements: the lexicon (*le mercure est un élément chimique*), the punctuation (the comma in *le Mercure, un fleuve du sud de l'Italie*), lexical inclusion (*appareil de mesure*, implying that *appareil de mesure* is an *appareil*), layout (disposition and typography) as used in enumerative structures (*la diode à vapeur de mercure est un appareil de mesure*, *la pile au mercure est un appareil de mesure*, etc.).

Depending on the language, the disambiguation pages are more or less structured. For instance, pages in English are very highly structured and contain essentially low-written text (noun-phrases and item lists). The French pages, in contrast, more readily mix the well-written text and the low-written text. So the French pages are particularly relevant for our experiment with a larger variety of ways to formulate hypernym relations.

The semantic resource that we use for building training examples is the BabelNet semantic network. BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network that connects concepts and named entities in a large network of semantic relations, made up of about 14 million entries. It has been automatically created by linking the encyclopedia Wikipedia to other resources, among which WordNet as source of lexical relations. In BabelNet, missing lexical entries in resource-poor languages have

[3] http://opennlp.apache.org/
[4] https://fr.wikipedia.org/wiki/Mercure



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.

Fig. 1. Excerpt from the French disambiguation page of the word *Mercure*

been filled with the help of statistical machine translation. As a consequence, the hypernym relation is well covered. Moreover, as long as BabelNet is derived from Wikipedia, the training knowledge base is derived from the training text, as recommended by the distant supervision approach.

We divided the set of disambiguation pages into two sets: a reference corpus composed of 20 pages, and a training corpus composed of the remaining pages (5904 pages). The two corpora have been pre-processed: (i) the plain text has been extracted from these pages, with the help of a WikiExtractor[5]; (ii) this plain text has been annotated with Part-Of-Speech and lemma (using TreeTagger); and (iii) it has been annotated with terms (including both single words and multiword expressions) corresponding to labels of concepts in BabelNet for the training corpus (resp. with the set of manually annotated relations for the reference corpus).

4.2. Features, examples and training models

4.2.1. Features

For each sentence in the training set, we extract a set of features from a window of size n. A sentence in the training text contains two terms tagged $Term1$ and $Term2$, that result from the annotation with the BabelNet concept labels. The window includes n tokens before $Term1$, $Term1$, the tokens between $Term1$ and $Term2$, $Term2$, n tokens after $Term2$. Only the features of the tokens belonging to this window are processed. We consider three kinds of features: those involving tokens, those involving sentences and those involving sentence windows. Currently, we focus on lexical and grammatical features, and some heuristics inspired by the work of Lin and colleagues[38], as they seem enough to provide good results. In fact, we decided no to use syntactic features because syntactic parsers provide poor results on low-written text. Furthermore, limiting the number of required NLP tools makes the approach easier to reproduce, especially for languages for which such tools are scarce. Table 1 presents the set of selected features.

4.2.2. Learning examples

Given that we only consider sentences that contain at least two terms denoting two concepts in BabelNet, the construction of training examples leans on the following assumptions:

[5] http://wiki.apertium.org/wiki/Wikipedia_Extractor

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

Table 1. Features set.

1. If two terms (in the same sentence) denote two concepts linked by a hypernym relation in BabelNet, a positive example will be built from the sentence window encompassing these two terms;
2. If no hypernym relation links the two BabelNet concepts whose labels occur in the same sentence, a negative example will be built from the sentence window encompassing these two terms.

Hence, for each sentence of each page of the training corpus, we randomly choose a pair of labeled terms[6] $Term1$ and $Term2$ occurring in that sentence. We then explore the BabelNet hierarchy to check whether $Term1$ and $Term2$ denote concepts linked by the hypernym relation. We empirically set a maximum length path of 3 levels in the hierarchy[39]. We give below an example of a feature vector, with a window size set to 3:

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

Matching the BabelNet list of terms lead to annotate the sentence with the 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 $Term2$. Tree Tagger parsing allows to replace the exact form of tokens by their part-of-speech followed by their lemma:

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

We finally compute distance features: for each token in the window, the feature is a pair of values representing the distance (in number of words) of this token respectively with $Term1$ and $Term2$. The last three features are the number of tokens between $Term1$ and $Term2$ (here 5); the number of tokens in the whole sentence (here 16); true or false depending on the presence (or absence) of a verbal form. This feature contributes to discriminate low-written text from well-written text.

```
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16 false
```

Here is the entire feature list for this example :

```
Terme1 KON/ou NOM/citron ADJ/vert PUN/, DET:ART/le Terme2 PRP:det/du NOM/limettier PUN/:
(0,6) (-1,5) (-2,4) (-3,3) (-4,2) (-5,1) (-6,0) (-7,-1) (-8,-2) (-9,-3) 5 16 false
```

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

4.2.3. Training models

From the whole set of 84169 training examples produced according to the process described above, 4792 examples are labeled as positive, and 79377 are labeled as negative. We randomly took 3000 positive examples and 3000

[6] In order to obtain reasonable computation time and to provide sets of examples of reasonable size, we do not compute all possible combinations of pairs of terms from a single sentence

negative examples. From these 6000 examples, 4000 were used as the set of training examples and 2000 as the test set (with a rate of 50% of positive examples, for both training and test sets).

We then produced different learning models (according to the supported features and with different sizes of sentence windows). We evaluated them in terms of precision, recall and F-measure, as shown in Table 2. We can observe that the best results were obtained for a window of size 3. This result can be explained on the one hand by the fact that with a window of size 1, we loose contextual information. On the other hand, a window of size 5 does not bring so much given that the corpus sentences are relatively short and low-written. Hence the contextual information obtained for that 5 word window is not discriminating.

Features : POS and Lemma			
	window size=1	window size=3	window size=5
Precision	0.67	0.69	0.69
Recall	0.72	0.78	0.76
F-measure	0.69	0.73	0.72

Table 2. Results for POS and Lemma features, for different window sizes.

From these results, we have trained another model using 3 as window size and all the features listed in Table 1. The results are reported in Table 3.

	Features : all features listed in Table 1 and window size=3
Precision	0.72
Recall	0.66
F-measure	0.69

Table 3. Results for all features of Table 1 and for a window size = 3.

Unlike what we expected, the model using only the POS and lemma with window of 3 (model called in the following `model_POSL`) outperforms also the model considering all the features listed in Table 1 and which we refer to as `model_AllFeatures` in the following, in terms of F-measure. This can be explained by the fact that for that kind of corpus, POS and lemma features seem to be discriminant. Nevertheless we decided to evaluate both models on the reference corpus.

5. Evaluation

The evaluation we present in this section aims at showing the performance (in terms of precision and recall) of the two learning models presented in the previous section (`model_POSL` and `model_AllFeatures`), when applied to a corpus with the same characteristics as the ones on which the models were trained. We also compare the results to two baselines, both based on lexico-syntactic patterns largely used in the literature to identify hypernym relations.

This evaluation concerns the reference corpus composed of 20 French disambiguation pages. Two annotators have annotated this corpus in a double blind process: 688 sentences contained 2 BabelNet terms have been annotated as true positives (hypernym relations) and 278 such sentences as true negatives (absence of hypernym relation). After the calculation of the annotation agreement (of 0.8) between annotators, all conflicts have been identified and resolved.

This corpus has been pre-processed as described in section 4.1. From sentences containing the manually annotated relations, we extracted feature vectors and submitted them to the two classifiers `model_POSL` and `model_AllFeatures`.

5.1. Baseline

As stated above, two baselines based on lexico-syntactic patterns have been used. A lexico-syntactic pattern is a regular expression composed of words, POS or semantic categories, and symbols aiming at identifying textual segments which match this expression. In the context of relation extraction, the pattern characterizes a set of linguistic forms whose interpretation is relatively stable and corresponds to a semantic relation between terms[40]. Patterns are

in fact very efficient, in particular in terms of precision, when they are adapted to the corpus. However, since their development is expensive, it is conventional to implement generic patterns such as those of Hearst[1]. We chose a more complete list of 30 patterns from the work of Jacques and Aussenac[41]. This set of patterns is our first baseline (called `Baseline 1` in the following).

In a second step, in order to better take into account the specificity of the corpus, we have defined ad-hoc patterns adapted to the low-written parts of the disambiguation pages[42]. The set of generic patterns together with the specific patterns is our second baseline[7] (called `Baseline 2` in the following).

5.2. Results and discussion

Table 4 shows the results of the two baselines and the two classifiers, in terms of precision, recall, F-measure and accuracy. We can observe that the `model_AllFeatures` model gives the best results, with a F-measure equal to that of `model_POSL` but with a better accuracy. As expected, in terms of precision, the generic patterns (`Baseline 1`) outperform all the other models, in detriment of recall. When using specific patterns (`Baseline 2`), recall is highly improved. However, they introduce some noise (false positives) and decrease precision.

	`Baseline 1`	`Baseline 2`	`Model_POSL`	`Model_AllFeatures`
Precision	0.96	0.81	0.68	0.71
Recall	0.04	0.46	0.66	0.63
F-measure	0.07	0.59	0.67	0.67
Accuracy	0.31	0.54	0.53	0.55

Table 4. Results for the two baselines and the two classifiers on the reference corpus.

Since the accuracy is substantially the same (except to `Baseline 1`), another way of looking at these results is given in Table 5, which presents the number of true positive hypernym relations per type of hypernym expression in the text, found by the baselines and by the classifiers. With the classifiers, we are able to identity the most recurrent forms of expression of the hypernym relation, namely those relations expressed with well-written text, as in "$Term1$ *is a* $Term2$", those expressed using layout and therefore expressed with low-written text, as in "$Term1$, $Term2$", and as well as those which can be identified thanks to the head-modifier method. While the learning approach is able to identify a larger variety of expression forms of the relation, outperforming the symbolic approach, the patterns introduced in the `Baseline 2` seem to fit well the corpus regularities.

	Relations expressed in well-written text	Relations expressed in low-written text	Relations expressed with head-modifiers	Total
`Baseline 1`	24	2	0	26
`Baseline 2`	23	294	0	317
`Model_POSL`	22	243	187	452
`Model_AllFeature`	31	261	139	431

Table 5. Results for the two baselines and the two classifiers for the different ways of expressing the hypernym relation.

The intersection of the results provided by these methods shows that 11 true positive hypernym relations were found both by `Baseline 1` and by the two classifiers, while 221 true positive hypernym relations were found both by `Baseline 2` and by the two classifiers. From a quantitative point of view, automatic learning identifies more examples than patterns, without any development cost, ensuring a systematic and less empirical approach. From a qualitative point of view, classifiers do not perform as well as patterns when relations are regularly expressed in the same way, as somehow expected, but they can identify more varied forms of relation expressions. For instance, patterns have not the ability to identify relations with head modifiers, while the classifiers are able to do so. The following examples show the expression of hypernym relations in complex sentences that could be correctly identified by the classifiers:

[7] A JAPE implementation of these two types of patterns is visible on the site:https://github.com/aghamnia/SemPediaPatterns

(1) <Louis Label, prêtre-missionnaire oblat> and <Louis Label, explorateur du Nouveau-Québec> in the sentence *Louis Babel, prêtre-missionnaire oblat et explorateur du Nouveau-Québec (1826-1912)*.

(2) <fontaine, robinet de cuivre> in the sentence *La fontaine a aussi désigné le "vaisseau de cuivre ou de quelque autre métal, où l'on garde de l'eau dans les maisons", et encore le robinet de cuivre par où coule l'eau d'une fontaine, ou le vin d'un tonneau, ou quelque autre liqueur que ce soit*.

In these examples, the relations are expressed within textual units using conjunction, as in example (1), or with $Term1$ and $Term2$ being relatively far from each other in the sentence, as in example (2). A pattern approach would require the definition of new patterns to fit these cases.

6. Conclusion and perspectives

This paper has presented a distant supervision learning approach for extracting hypernym relations from a corpus where hypernymy is expressed in a variety of forms. We composed a corpus from Wikipedia disambiguation pages, which are rich in hypernym relations expressed through textual definitions and named entities. We have mainly used lexical and grammatical features. Even with this reduced set, we could observe that the learning approach is able to extract regularities from the corpus. It was able to correctly identify different ways of expressing relations, including a set of those that could be identified by patterns or head-modifiers, for instance.

We have evaluated learning and patterns independently. However their combination seems to be a good strategy. In fact, combining learning with a broad set of patterns, best suited to the corpus, could ensure the best results (at the cost of developing such patterns). In the case of very regular pages with singularities (such as disambiguation pages), the development of ad-hoc patterns is immediate and justified.

As future work, we plan to train a model on the whole set of Wikipedia pages. We also intend to investigate additional features such as semantic, distributional or lay-out features. Moreover, we plan to study how patterns and learning approaches can be better combined.

Acknowledgement

The Sempedia Project is funded by the French region Occitanie-Mediterranée-Pyrénées from 2015 to 2018. This work also contributed to the SparkinData project funded by the French FUI program from 2015 to 2017.

References

1. Hearst, M.A.. Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of the 14th conference on Computational linguistics*. Association for Computational Linguistics; 1992, p. 539–545.
2. Lenci, A., Benotto, G.. 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*. Association for Computational Linguistics; 2012, p. 75–79.
3. Pantel, P., Pennacchiotti, M.. Automatically harvesting and ontologizing semantic relations. *Ontology learning and population: Bridging the gap between text and knowledge* 2008;:171–198.
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.. Open information extraction from the web. In: *IJCAI*; vol. 7. 2007, p. 2670–2676.
5. Malaisé, V., Zweigenbaum, P., Bachimont, B.. Detecting semantic relations between terms in definitions. In: Ananadiou, S., Zweigenbaum, P., editors. *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*. Geneva, Switzerland: COLING; 2004, p. 55–62.
6. Fauconnier, J.P., Kamel, M.. Discovering Hypernymy Relations using Text Layout (regular paper). In: *Joint Conference on Lexical and Computational Semantics (SEM), Denver, Colorado, 04/06/2015-05/06/2015*. Association for Computational Linguistics (ACL); 2015, p. 249–258.
7. Morsey, M., Lehmann, J., Auer, S., Stadler, C., Hellmann, S.. Dbpedia and the live extraction of structured data from wikipedia. *Program* 2012;**46**(2):157–181.
8. 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*. 2009, p. 1003–1011.
9. Navigli, R., Ponzetto, S.P.. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 2012;**193**:217–250.

10. Séguéla, P., Aussenac-Gilles, N.. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In: *Conférence ingénierie des connaissances*. 1999, p. 79–88.
11. Morin, E., Jacquemin, C.. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities* 2004;**38**(4):363–396.
12. Berland, M., Charniak, E.. Finding parts in very large corpora. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics; 1999, p. 57–64.
13. Snow, R., Jurafsky, D., Ng, A.Y.. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17* 2004;.
14. 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*. Association for Computational Linguistics; 2005, p. 724–731.
15. Brin, S.. Extracting patterns and relations from the world wide web. In: *International Workshop on The World Wide Web and Databases*. Springer; 1998, p. 172–183.
16. Agichtein, E., Gravano, L.. Snowball: Extracting relations from large plain-text collections. In: *Proceedings of the fifth ACM conference on Digital libraries*. ACM; 2000, p. 85–94.
17. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., et al. Web-scale information extraction in knowitall:(preliminary results). In: *Proceedings of the 13th international conference on World Wide Web*. ACM; 2004, p. 100–110.
18. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.. 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*. Association for Computational Linguistics; 2007, p. 25–26.
19. Fader, A., Soderland, S., Etzioni, O.. Identifying relations for open information extraction. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2011, p. 1535–1545.
20. Fabre, C., Hathout, N., Ho-Dac, L.M., Morlane-Hondère, F., Muller, P., Sajous, F., et al. 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; 2014, p. 196–205.
21. Yap, W., Baldwin, T.. Experiments on pattern-based relation learning. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM; 2009, p. 1657–1660.
22. Granada, R.L.. *Evaluation of methods for taxonomic relation extraction from text*. Ph.D. thesis; Pontifícia Universidade Católica do Rio Grande do Sul; 2015.
23. Kamel, M., Aussenac-Gilles, N.. How can document structure improve ontology learning? In: *Workshop on Semantic Annotation and Knowledge Markup collocated with K-CAP*. 2009, .
24. O'Connor, M.J., Das, A.. Acquiring owl ontologies from xml documents. In: *Proceedings of the Sixth International Conference on Knowledge Capture*. New York, NY, USA: ACM. ISBN 978-1-4503-0396-5; 2011, p. 17–24.
25. Sumida, A., Torisawa, K.. Hacking wikipedia for hyponymy relation acquisition. In: *IJCNLP*; vol. 8. Citeseer; 2008, p. 883–888.
26. Jannink, J.. Thesaurus entry extraction from an on-line dictionary. In: *Proceedings of Fusion*; vol. 99. Citeseer; 1999, .
27. Chernov, S., Iofciu, T., Nejdl, W., Zhou, X.. Extracting semantics relationships between wikipedia categories. *SemWiki* 2006;**206**.
28. Suchanek, F.M., Kasneci, G., Weikum, G.. Yago: a core of semantic knowledge. In: *Proceedings of the 16th international conference on World Wide Web*. ACM; 2007, p. 697–706.
29. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.. Dbpedia: A nucleus for a web of open data. In: *The semantic web*. Springer; 2007, p. 722–735.
30. 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*. 2007, p. 698–707.
31. Flati, T., Vannella, D., Pasini, T., Navigli, R.. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artificial Intelligence* 2016; **241**(Complete):66–102. doi:10.1016/j.artint.2016.08.004.
32. Rodriguez-Ferreira, T., Rabadan, A., Hervas, R., Diaz, A.. Improving information extraction from wikipedia texts using basic english. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA); 2016, .
33. 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'07)*. Prague, Czech Republic; 2007, .
34. Riedel, S., Yao, L., McCallum, A.. Modeling relations and their mentions without labeled text. In: *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*; ECML PKDD'10. Berlin, Heidelberg: Springer-Verlag; 2010, p. 148–163.
35. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.. A maximum entropy approach to natural language processing. *Computational linguistics* 1996; **22**(1):39–71.
36. Jaynes, E.. Information theory and statistical mechanics. *Physical review* 1957;**106**(4):620.
37. Ratnaparkhi, A.. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis; University of Pennsylvania; 1998.
38. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.. Neural relation extraction with selective attention over instances. In: *ACL*. 2016, .
39. Kamel, M., Trojahn, C.. 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 (Proceedings of the 27th French Knowledge Engineering Conference), Montpellier, France, June 6-10, 2016*. 2016, p. 111–122.
40. Rebeyrolle, J., Tanguy, L.. Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire* 2000;**25**:153–174. URL: `https://halshs.archives-ouvertes.fr/halshs-01322256`.
41. Jacques, M.P., Aussenac-Gilles, N.. Variabilit des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues, Non Thmatique* 2006;**47**(1):(en ligne).
42. Ghamnia, A.. Extraction de relations dhyperonymie partir de wikipdia. In: *Actes de la confrence conjointe JEP-TALN-RECITAL*. 2016, .