MIsA: Multilingual "IsA" Extraction from Corpora

Stefano Faralli[1]**, Els Lefever**[2]**, Simone Paolo Ponzetto**[1]
[1]Data and Web Science, University of Mannheim, University of Mannheim
{stefano, simone}@informatik.uni-mannheim.de
[2]LT3, Language and Translation Technology Team, Ghent University, Belgium
els.lefever@ugent.be

Abstract

In this paper we present Multilingual IsA (MIsA), which is a collection of hypernymy relations in five languages (i.e., English, Spanish, French, Italian and Dutch) extracted from the corresponding full Wikipedia corpus. For each language, we first established a set of existing (viz. found in literature) or newly defined lexico-syntactic patterns. Similarly to WebIsADb, the resulting resource contains hypernymy relations represented as "tuples", as well as additional information such as provenance (Gil and Groth, 2011), context of the extraction, etc. To measure the precision of the patterns, we performed a manual assessment of the quality of the extracted relations and an error analysis. In addition, we release the software developed for the extraction of the hypernym tuples.

Keywords: multilinguality, Hearst patterns, hypernym extraction, framework.

1. Introduction

Hypernymy relations represent the relationship between a generic term (hypernym) and a specific instance of it (hyponym). These relations play a key role for many Natural Language Processing (NLP) tasks, e.g. ontology learning, automatically building or extending knowledge bases, or word sense disambiguation and induction. In fact, hypernymy relations may provide the basis for the construction of more complex structures such as taxonomies, or be used as effective background knowledge for many word understanding applications.

In the past, many different methods have been developed for hypernym extraction, ranging from simple lexical patterns (Hearst, 1992; Oakes, 2005) to statistical and machine learning techniques (Dolan et al., 1993; Caraballo, 1999; Agirre et al., 2000; Ritter et al., 2009), to name a few.

Snow et al. (2004) first search sentences that contain two terms that are known to be in a taxonomic relation (term pairs are taken from WordNet (Miller et al., 1990)), then parse the sentences, and automatically learn patterns from the parse trees. Finally, they train a hypernym classifier based on these features. Lexico-syntactic patterns are generated for each sentence relating a term to its hypernym, and a dependency parser is used to represent them.

For the ontology learning task, Velardi et al. (2013) induce taxonomies from scratch by extracting hypernyms from a domain corpus and the Web. Definitional sentences such as *"lion is a dangerous animal"* (where *"animal"* is the hypernym of *"lion"*) are recognized by the Word Class Lattices classifier (Navigli and Velardi, 2010) trained on a large set of Wikipedia definitions.

Kozareva and Hovy (2010) induce a taxonomy using a particular kind of Hearst-like (Hearst, 1992) lexico-syntactic patterns, i.e. so-called Doubly Anchored Patterns (DAP). The hypernymy relation extraction consists of two phases. First, the authors bootstrap the terminology harvesting with DAP of the kind *"animals such as lions and *"*, so it is possible to discover new terms such as *"cats"*. Next, for each pair of terms in the discovered terminology, e.g. *("lions","cats")*, they automatically create a DAP^{-1} of the kind *"* such as lions and cats"* and discover new hypernyms (e.g. *"felines"*).

The above mentioned works focus on domain-specific hypernymy relations extraction and due to their need of domain constraints - a specific defined term in (Navigli and Velardi, 2010) or a seed pair (Kozareva and Hovy, 2010) - they can not be used to collect the whole set of hyponym-hypernym pairs from a large scale corpus such as the Web.

For Microsoft's *Probase* (Song et al., 2011), albeit not freely accessible, the authors used Hearst-like lexico-syntactic patterns to extract hypernymy relations from 1.68 billion web pages in Microsoft Bing's web corpus, instead of focusing on domain specific hypernymy relations. Probase's main purpose was to create a universal taxonomy containing more than 2.7 million concepts. To this end, the methods underlying Probase are able to extract approximately 25 million pairs.

The interest in the hypernymy extraction task is also illustrated by two shared tasks organised within the SemEval framework: TExEval (Taxonomy Extraction Evaluation) focused on finding hyponym-hypernym relations between a list of domain-specific English terms and subsequent taxonomy construction (Bordea et al., 2015), whereas TExEval-2 introduced a multilingual setting for this task, covering four different languages (English, Dutch, Italian and French) from domains as diverse as environment, food and science (Bordea et al., 2016).

Our MIsA is an extension of the WebIsADb framework (Seitner et al., 2016) - a publicly available database with more than 400 million English hypernymy relations extracted from the CommonCrawl web corpus - where:

1. we investigate and evaluate the performance of a collection of existing and new lexico-syntactic patterns for five languages of interest (i.e., English (EN), Spanish (ES), French (FR), Italian (IT), Dutch (NL));

2. we release a new standalone, language-independent and easy to adapt/configure extractor, which is ready to ex-

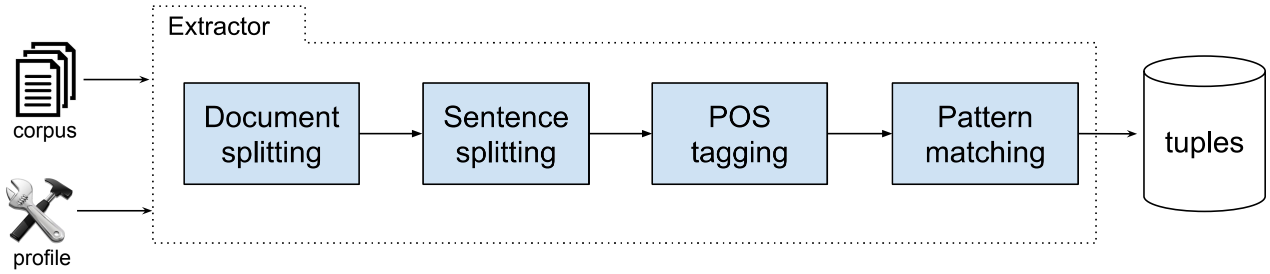

Figure 1: Pipeline for the extraction of language specific "IsA" tuples from a corpus.

Table 1: Statistics of the selected multilingual patterns.

	# patterns	excerpt
EN	74	NP$_t$, is a NP$_c$
		NP$_c$, such as NP$_t$
		NP$_t$, and any other NP$_c$
ES	33	NP$_t$, es un NP$_c$
		la NP$_t$ es una de las NP$_c$
		NP$_t$, y otros NP$_c$
FR	60	NP$_t$, est un NP$_c$
		NP$_c$, comme par exemple NP$_t$
		NP$_t$, et autres NP$_c$
IT	78	NP$_t$, è un NP$_c$
		NP$_t$ in confronto ad altre NP$_c$
		NP$_t$, e altri NP$_c$
NL	14	NP$_t$, is een NP$_c$
		NP$_c$ is een soort van NP$_t$
		NP$_t$ en veel andere NP$_c$

tract hypernymy relations from text corpora in five languages and which can be easily extended to work with other languages;

3. we release more than 32 million hypernymy relations extracted from the corresponding Wikipedia corpus for the five languages of interest.

Both datasets and tools can be downloaded from `http://web.informatik.uni-mannheim.de/misa/`

2. Multilingual patterns

As introduced in Section 1., we focused our efforts on five languages of interest (i.e., English, Spanish, French, Italian and Dutch). In Table 1 we show some statistics about our collections of Hearst-like (Hearst, 1992) lexico-syntactic patterns (the full list of selected patterns is available at `http://web.informatik.uni-mannheim.de/misa/`).

Lexico-syntactic patterns are defined as regular expressions where NP_t and NP_c are special expressions to match a noun-phrase for the definiendum and the hypernym respectively. In order to identify a noun phrase NP, similarly to WebIsADb (Seitner et al., 2016) we apply part-of-speech tagging to obtain the mapping of words to grammatical categories and identify a NP on a disjunctive/conjunctive sequence of accepted part-of-speech. With this definition of NP we are able to intercept sequences of concepts. For example, given the English pattern "NP$_c$, such as NP$_t$" we are able to match sentences like "pure-bred dogs *such as* a bulldog or pug" (where $NP_c =$ "pure-bred dogs" and $NP_t =$ "bulldog"|"pug" is a sequence of concepts) and to produce multiple hypernymy relations from a single match (e.g., (bulldog, pure-bred dogs) and (pug, pure-bred dogs)).

Some patterns are directly selected or translated from literature works, such as: i) Ponzetto and Strube (2011), where *isa* patterns were used to induce a taxonomy from Wikipedia; ii) Orna-Montesinos (2011), where patterns for the term “building” were extracted on a set of specialized textbooks in the field of construction engineering; iii) Klaussner and Zhekova (2011) where the authors extract *IsA* relations from selected Wikipedia pages and iv) research describing lexico-syntactic patterns for languages other than English, such as Lefever et al. (2014) for Dutch, Séguéla (2001) for French and Galicia-Haro and Gelbukh (2014) and Ortega-Mendoza et al. (2007) for Spanish. The remaining are brand-new experimental patterns, whose selection is dictated mainly from the experience of experts in the field of NLP. In Section 5., we provide a manual assessment of the quality of the most productive patterns that were selected for the five considered languages.

3. Extraction Framework

In Figure 1 we show a diagram representing the pipeline of our extraction framework. Our pipeline input includes:

- a *corpus* in the form of a collection of flat text (UTF-8) documents written in a specific language of interest;
- a language *profile* including the following language-specific information:
  - *patterns:* the definition of lexical-syntactic patterns (as introduced in Section 2.) to be matched on the input corpus;
  - *abbreviations:* a list of abbreviations (e.g., the English abbreviation "Prof." for "Professor");
  - *pronouns:* lists of *demonstratives*, *personals*, *possessives* and *interrogative* pronouns;
  - *conjunctions:* to identify sequences of concepts (e.g., the English "or", "and").

Table 2: The total number of tuples extracted for each of the five selected languages.

	tuples	example
EN	23,386,043	. . . [*understory plants*]$_c$ such as [*bushes*]$_t$ and [*vines*]$_t$. . . . . . [*shrubs*]$_c$ including [*mountain laurel*]$_t$ and [*rhododendron*]$_t$. . . [*Mdawrush*]$_t$ is a [*municipality*]$_c$ . . .
ES	3,649,166	. . . [*Carnavalón*]$_t$, el cuales una [*ceremonia*]$_c$ acompañada de mùsica, . . . [*Paella*]$_t$ es una [*receta de cocina*]$_c$
FR	2,390,867	[*Saint-Claude*$_t$] est un [*village*]$_c$. . . [*Saint-Mars-sur-Colmont*$_t$] est une [*commune française*]$_c$. . . . . . [*Ponte alla Vittoria*$_t$] est un des [*ponts*]$_c$. . .
IT	774,964	La [*salumeria*]$_t$ è un [*negozio*]$_c$. . . [*Roma*]$_t$ è la [*capitale*]$_t$ della Repubblica Italiana,. . .
NL	1,844,644	[*Framboise Boon*]$_t$ is een [*fruitbier*]$_c$. . . [*vlees*]$_c$ zoals [*biefstuk*]$_t$, [*kip*]$_t$, [*varkenskarbonades*]$_t$, [*schapenvlees*]$_t$ en [*lamskoteletten*]$_t$. . .

where *abbreviations*, *pronouns* and *conjunctions* are devoted to language specific NP identification.

The extraction process is then divided into four main steps:

1. *Document splitting:* in the initial phase the corpus is structured as a collection of indexed documents and titles are collected to later identify all the extraction's provenances;

2. *Sentence splitting:* since the context of the extraction is a single sentence, we split each document in separated one-line sentences;

3. *POS tagging:* each sentence is processed with a POS tagger (we use the Stanford POS-tagger (Toutanova et al., 2003) for the EN, FR and ES corpora and TreeTagger (Schmid, 1994) for the IT and NL corpora) to allow identification of NPs in the next step;

4. *Pattern matching:* in this final step we find all matches between the lexico-syntactic patterns and the POS-tagged input sentence.

The output of our pipeline consists of a collection of tuples representing each pattern match, including (1) the pattern that is matched, (2) a pair of concept sequences (T, C) where T is the sequence of definiendum NPs and C is the sequence of hypernymy NPs, (3) the provenance (i.e., the corpus and document identifier), (4) the contextual matching sentence and (5) the POS-tag sequence for the sentence.

4. Multilingual Resource

We applied the pipeline described in Section 3. on five language specific Wikipedia dumps (latest available dumps accessed on 2 May 2017) with the corresponding language profiles including the patterns described in Section 2.. In Table 2 we show, for each language, the resulting number of generated tuples.

5. Evaluation

Similar to the WebIsADb, our aim is to provide with MIsA both a tool and a resource to favour investigations and applications with lexico-syntatic patterns. To this end, we provide in Section 5. an assessment of the precision of the extracted hypernymy pairs and a qualitative error analysis.

5.1. Manual assessment of precision

We show the results of our manual assessment of the quality of the extracted hypernymy relations for the five most productive patterns per language (30 patterns in total). For each pattern, a random sample of 100 extracted hypernym tuples was manually verified by the annotators, who assigned one of the following three labels to each matched hypernym pattern:

1. *Correct*: correctly extracted hypernym tuple.

2. *Partially correct*: the extracted hypernym tuple is not complete (missing hyponyms, part of the instance/class is missing, e.g. *Operation Little Switch was an exchange of sick and wounded prisoners* resulting in (Operation Little Switch, exchange)) or is too context-dependent or vague (e.g. *John Laurence is a friend* resulting in (John Laurence, friend)).

3. *Not Correct*: wrongly extracted hypernym tuple.

During the annotation process, the annotators also had access to the accompanying information as described in Section 3. Table 3 shows an example of the input the annotators were provided with, containing amongst others the hypernym tuple, the original input sentence containing the hypernym pair, the POS-tag sequence of the class and instance terms, etc.

We show in Table 4 the resulting estimated precision for the most productive patterns across the five languages of interest. As expected, when combining the correct and the partially correct matches, we observe similar pattern behaviors as in the WebIsADb (Seitner et al., 2016).

5.2. Error Analysis

The aim of this work is a robust hypernym extraction system, which can be easily deployed on very large (web) corpora. We implemented patterns for 5 languages, aiming for a high recall, sometimes at the cost of precision. As we make the code freely available to the community, researchers can easily adapt the code to add patterns or improve the precision of the current implementation.

Table 3: Example of the input provided for the manual labeling of the pattern quality.

Information	example
ID	Mastertapes.txt
Pattern	EN_p8a: NP$_t$, is a NP$_c$
NP$_t$	Mastertapes
NP$_t$ POS	NN
NP$_c$	BBC Radio 4 programme
NP$_c$ POS	NN NN CD NN
Input	Mastertapes is a BBC Radio 4 programme, presented by John Wilson, which discusses the making of significant rock albums.
POS	Mastertapes/NN, is/VBZ, a/DT, BBC/NN, Radio/NN, 4/CD, programme/NN, presented/VBN, by/IN, John/NN, Wilson,/NN, which/WDT, discusses/VBZ, the/DT, making/NN, of/IN, significant/JJ, rock/NN, albums/IN

Table 4: Number of matches and estimated precision for the most productive patterns across languages

	Pattern	#matches	Precision	Partial match
EN	NP$_t$, is a NP$_c$	1,855,931	48%	25%
	NP$_c$, such as NP$_t$	90,7430	50%	21%
	NP$_c$, including NP$_t$	894,625	34%	15%
	NP$_t$, was ((a)|(an)) NP$_c$	641,385	31%	25%
	examples of NP$_c$, are NP$_t$	501,122	6%	0%
ES	NP$_t$ es una NP$_c$	1,413,307	59%	9%
	NP$_t$ era una NP$_c$	311,044	34%	12%
	NP$_t$ era un NP$_c$	311,010	29%	15%
	NP$_t$ eran una NP$_c$	80,163	27%	25%
	NP$_t$ eran un NP$_c$	79,891	20%	35%
FR	NP$_t$ est un NP$_c$	600,613	52%	18%
	NP$_t$ est une NP$_c$	444,456	38%	17%
	NP$_t$ est ((un)|(une)) des NP$_c$	168,006	18%	13%
	NP$_t$ nommé NP$_c$	104,938	10%	6%
	NP$_t$ sont ((le)|(la)|(les)) NP$_c$	82,141	1%	1%
IT	NP$_t$ è un NP$_c$	331,352	65%	24%
	NP$_t$ è una NP$_c$	147,759	62%	24%
	NP$_c$ in particolare NP$_t$	38,107	24%	18%
	NP$_t$ (e|o|(ed)|(oppure)) altri NP$_c$	22,029	59%	9%
	NP$_t$ era un NP$_c$	21,987	48%	15%
NL	NP$_t$, is een NP$_c$	1,182,272	61%	13%
	NP$_c$, (zo)?als NP$_t$	634,383	10%	12%
	NP$_c$, (en|of) (veel)? ander(e)? NP$_t$	169,116	9%	3%
	NP$_t$, zijn een NP$_c$	24545	4%	14%
	NP$_c$, (inclusief|specifiek) NP$_t$	3,258	13%	12%

A manual analysis of the extracted hypernym tuples revealed several possibilities to improve the current implementation. One obvious way to increase the precision of the system is to add additional syntactic constraints to the regular expressions.

Two recurrent phenomena causing overgeneration of the patterns appeared to be:

1. hypernyms/hyponyms extracted from **prepositional phrases** (e.g. *Houston has had notable sports teams in its history, including Phi Slama Jama*, where (Phi Slama Jama, history) is extracted by the pattern “NP$_c$, including NP$_t$");

2. **wrong part-of-speech** tagging (e.g. *Abuse of positive leverage can also lead to coercion, including bribery and blackmail*, where *coercion* is tagged as a verb, which prevents the hypernym pattern “NP$_c$, including NP$_t$" from matching the input sentence).

Finally, it is also important to mention that the current implementation of the system does not take into account agreement between the hypernym and hyponym terms. As an example, we can cite the tuple (*Basilica San Marco*, *churches*), where the instance is a singular entity, whereas the class refers to a plural noun. A simple solution for these agreement problems could be the lemmatisation of all extracted hypernym and hyponym terms.

6. Conclusion

We presented *MIsA*, a multilingual collection of hypernymy relations extracted from five language specific Wikipedia dumps. The resource is created by means of language-specific sets of Hearst-like patterns that were collected from literature or dictated by experimental needs. In fact, the aims of this research are: i) to experiment with multilingual Hearst-like patterns on large corpora; ii) to provide an evaluation of both the quality and the limits of lexico-syntactic pattern approaches; iii) to release a versatile tool to let other researchers extend (with minimal effort) our case study to other languages or to different selections of language specific patterns. Future work will include the extension of the multilingual setting to larger corpora (e.g., Web-scale corpora) and to other languages.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft within the JOIN-T project (research grant PO 1900/1-1).

7. Bibliographical References

Agirre, E., Ansa, O., Hovy, E. H., and Martínez, D. (2000). Enriching very large ontologies using the WWW. In *Proceedings of the ECAI Workshop on Ontology Learning*.

Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. (2015). Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of SemEval-15*.

Bordea, G., Lefever, E., and Buitelaar, P. (2016). Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of SemEval-16*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of ACL-99*, pages 120–126.

Dolan, W., Vanderwende, L., and Richardson, S. D. (1993). Automatically deriving structured knowledge bases from on-line dictionaries. In *Proceedings of PACLING-93*, pages 5–14.

Galicia-Haro, S. and Gelbukh, A. (2014). Extraction of semantic relations from opinion reviews in spanish. In *Proceedings of MICAI-14*, pages 175–190.

Gil, Y. and Groth, P. T. (2011). Using provenance in the semantic web. *J. Web Sem.*, 9(2):147–148.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*, pages 539–545.

Klaussner, C. and Zhekova, D. (2011). Pattern-based ontology construction from selected wikipedia pages. In *Proceedings of the RANLP-11 SRW*.

Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of EMNLP-10*, pages 1110–1118.

Lefever, E., Van de Kauter, M., and Hoste, V. (2014). Hypoterm detection of hypernym relations between domain-specific terms in Dutch and English. *Terminology*, 20(2):250–278.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., and Miller, K. (1990). WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of ACL-10*, pages 1318–1327.

Oakes, M. P. (2005). Using Hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. In *Proceedings of the RANLP Text Mining Workshop*, pages 63–67.

Orna-Montesinos, C. (2011). Words and patterns: lexico-grammatical patterns and semantic relations in domain-specific discourses. *Revista Alicantina de Estudios Ingleses*, 24:213–233.

Ortega-Mendoza, R., Villaseñor Pineda, L., and Montes-y Gómez, M. (2007). Using lexical patterns for extracting hyponyms from the web. In *Proceedings of MICAI-07*, pages 904–911.

Ponzetto, S. P. and Strube, M. (2011). Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10):1737–1756.

Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 88–93.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

Séguéla, P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Ph.D. thesis, Université Toulouse III Paul Sabatier.

Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., and Ponzetto, S. P. (2016). A large database of hypernymy relations extracted from the web. In *Proceedings of LREC-16*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS-04*, pages 1297–1304.

Song, Y., Wang, H., Wang, Z., Li, H., and Chen, W. (2011). Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of IJCAI-11*, pages 2330–2336.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-03*, pages 173–180.

Velardi, P., Faralli, S., and Navigli, R. (2013). OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.