Lexico-Syntactic Patterns for Automatic Ontology Building

Carmen Klaussner
University of Nancy 2
carmen@wordsmith.de

Desislava Zhekova
University of Bremen
zhekova@uni-bremen.de

Abstract

In this paper, we evaluate different lexico-syntactic patterns in regard to their usefulness for ontology building. Each pattern is analysed individually to determine its respective probability to return the hyponymy relation. We also create different ontologies according to this accuracy criteria to show how it influences the resulting ontology. Using patterns with a success rate over 80% leads to an approximate accuracy of 77% in the final ontology.

1 Introduction

Computers have become increasingly important in the communication and usage of information. Therefore, also the way in which information is prepared for processing by computers has gained in interest, since machines do not possess human-comparable skills in regard to Natural Language Processing, when, for instance, solving issues of ambiguity (Lacy, 2005). Machines need knowledge bases that offer clearly structured and meaningful representation of information. However, information is not static, but instead constantly changing. This makes hand-crafted, reliable knowledge bases, such as *WordNet*[1] not feasible, since its constant extensions to ensure a continuing coverage would result in exceedingly-high costs. Automatic ontology building is one approach to address this issue. An ontology is a type of knowledge representation that is understandable to both humans and computers. It is populated by definitions or facts that are organised into hierarchies. These thereby model relationships of and dependencies between entities in the world. Automatic ontology building can be realised in different ways. One approach is pattern-based extraction of definition relations, which are then converted into the respective ontology representation. Pattern-based extraction has shown quite reasonable success rates, while it is easy to implement and can be applied to unrestricted text (Hearst, 1998). Although using lexico-syntactic patterns for ontology building is reasonably successful, ambiguous patterns, which return correct as well as incorrect results, remain problematic since they can lead to an overall decrease in accuracy for the whole ontology.

In the present paper, we assess various lexico-syntactic patterns that model the semantic relation of hyponymy in order to identify those, which are both frequent and reliable to return this relation. These patterns have been classified as successful in connection with other knowledge sources, whereas we aim to measure their reliability with *Wikipedia*. Our hypothesis is, that the usage of reliable lexico-syntactic patterns indicative of hyponymy, return relations that can create useful, widely-applicable ontologies. The latter are suitable as knowledge bases in many computational linguistic applications (e.g. Machine Translation, Information Extraction, Text Generation, etc).

Thus, section 2 gives a short overview of related work projects and approaches. In section 3, we introduce the system that we use for the automatic ontology building – the *Ontology creator (Oc)*[2]. We also describe how patterns are employed, while further, in section 4, we evaluate the different lexico-syntactic patterns in regard to their accuracy and describe the most common issues that we observed during our experiments. We create different ontolgies in order to effectively investigate to what extent using successful/unsuccessful patterns influences the overall accuracy of the final outcome. In section 5 we conclude and suggest further approaches for the advancement of our work.

[1] http://wordnet.princeton.edu/

[2] http://sourceforge.net/projects/ontocreation/

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 109–114,
Hissar, Bulgaria, 13 September 2011.

2 Related Work

There has been considerable work in regard to pattern-based extraction of information. Hearst (1992), for instance, identified a method for discovering new lexico-syntactic patterns. This entails searching corpora for specific terms that are connected through a semantic relation and deriving possible patterns from the results. If they prove to successfully return the same relation, these patterns can be applied domain-independently in order to identify and extract definitions. Lexico-syntactic patterns can model various semantic relations, although hyponymy seems to yield the most accurate results (Hearst, 1992). Moreover, they have the advantage of a frequent occurrence across many different text genres, and a reasonable overall accuracy even with little or no pre-encoded knowledge (Hearst, 1998). Mititelu (2006) also pursued the same aim and applied a slightly different method for discovering patterns, while working with English corpora. For some patterns, the subsequent success rates were as high as 100% (Mititelu, 2008).

Another approach very similar to ours is the one presented by Maynard et al. (2009). The authors also use lexico-syntactic patterns for the automatic creation of ontologies, but since they do not restrict their set of extracted relations only to hyponymy, the final ontology hardly reaches 50% precision. The authors conclude that the achieved results are very promising, however, they see the need for further improvement and refinement of the used lexico-syntactic patterns.

3 Pattern-Based Ontology Construction

The *Oc*, which was conceived for automatic ontology building, consists of different parts, that are presented in the following section. Section 3.1 introduces ontologies and the hyponymy relation, that forms the basis for the lexico-syntactic patterns. We show how different definition types, that were returned by the patterns, are transformed into an ontology representation using the web ontology language *OWL*. Section 3.2 describes the outer modules that were integrated into the *Oc* to obtain a knowledge source for the definition search.

3.1 Patterns in Ontology Building

In the context of computer and information sciences, an ontology is a machine-readable collection of terms and is used in knowledge sharing and reuse. Ontologies can encode models of the world, that is: objects, concepts, entities and the relationships that hold between them. Ontologies can be constructed on a textual basis and encoded into files using ontology languages. *OWL*[3] is one of the languages that can be used for this purpose. Relationships between entities in *OWL* exist between superclasses and subclasses or superclasses and individuals/members. Classes may have subclasses, which introduce more specific concepts than their superclass, or members/instantiations of a particular class concept. Their relation is generally one of hyponymy (in the sense that: If NP_i is a (kind of) NP_0, then for $1 \leq i$, hyponym (NP_i, NP_0) (Hearst, 1998)) or the *IS-A* link. This represents one of the most basic types of conceptual relations carrying with it the notion of an explicit taxonomic hierarchy, which allows all members of a particular superclass to inherit the properties of that class (Brachman, 1983). In *OWL*, these attributes of class members are introduced by the property relation and can be restricted through the superclass.

Lexico-syntactic patterns are suitable for automatic ontology building, since they model semantic relations. These display exactly the kind of relation between their parts that makes them easily translatable into an ontology representation. The lexico-syntactic pattern in (1) (Hearst, 1992) corresponds to the classic hyponymy relation:

(1) If $(NP_0\ such\ as\ NP_1, NP_2..., (and \mid or) NP_n)$
$for\ all\ NP_i,\ 1 \leq i \leq n, hyponym(NP_i, NP_0)$

The pattern specification as in (1) is able to identify and match sentences, as for example: *"The other major European powers, such as the UK, still had high fertility rates..."* Consequently, a lexico-syntactic pattern is a reoccuring environment that is indicative of a certain relationship between two or more entities. Having identified a lexico-syntactic pattern for a particular relation, it can usually be applied to unrestricted text and across different genres. When these relations are then transferred into *OWL*, there are different issues to be considered. First of all, there is the decision of whether to make a new entity an individual rather than a class. In this context, where there are only general indications of how the results will look like, the processing approach has to

[3] http://www.w3.org/TR/owl-ref/

NightTerrors
⊓
ParasomniacDisorder
⊓
(Disorder ⊓ hasCharacterisitic some Parasomniac)
⊓
Disorder

Figure 1: Simple subclass example in OWL

be one that is likely to be suitable in most cases.

All NP_0s become superclasses, since all of them will have either members or subclasses and should therefore constitute a class. A NP_{1+i}, on the other hand, will only be an individual, when all its substrings have been classified as proper nouns by the parser, otherwise it will be a subclass. Modifiers are generally set to become subclasses of the predefined `characteristicValues` class and linked to its class through the `hasCharacteristic` property. Modifiers to both NP_0 and NP_{1+i} also determine the number of superclass/subclass levels that are created. For example, if we consider the match *"....night terrors other than parasomniac disorders ..."* (leading to the relation: *hyponym("night terrors"-NP_1, "parasomniac disorders"-NP_0))*, where a modifier of NP_0 is present (as visualised in figure 1). First a general class `Disorder` is created. Through an intersection with `hasCharacteristic some Parasomniac` and `Disorder`, it will be indirect superclass to `NightTerrors`. It is generally assumed, that nouns that are modified by some adjective would otherwise constitute an own concept and will only be more specific through this addition. For two joined nouns, we could not make the same generalisation, since not all of them share this construction, where one concept modifies another and each convey a separate concept.

The conversion of NP_0s featuring a head with a complement leads to multiple problematic cases, such as varying scope and irregularities in processing. Yet, it is not our goal to discuss them here, since they are presented in more detail in (Klaussner and Zhekova, 2011).

The *Oc* uses the *OWL DL dialect*, which supports reasoning and thus inference of new facts from existing ones. It also allows to determine whether an ontology is consistent (inconsistency is then the case when an individual is a member of two mutually disjoint classes, e.g. an instance that is young and old at the same time). Although *OWL* allows to mark this mutual distinctness of members or classes, we cannot make all classes or individuals of a match mutually distinct/disjoint, since two names can often refer to the same individual. Only patterns that specifically indicate different subtypes can be processed in this way.

3.2 Knowledge Resource

For the purpose of testing our patterns and the later building of ontologies, we used articles obtained from *Wikipedia*, which has the advantage of being a regularly-updated knowledge resource, that contains articles on a wide variety of topics, although without an explicit hierarchy. The articles were extracted by a webcrawler, which is given a specific search term, which it will then use to further collect pages that have a referring link to it. Building a domain-specific ontology on possibly only one area of knowledge, requires a collection of articles of which as many as possible will be topically-interlinked. For this project, we chose the open-source webcrawler *JSpider*[4], which is a highly configurable Web Spider engine. It allows to limit the search to only one website, set the depth into its structure as well as the MIME type and the number of to be fetched resources per site. These features are all important to keep the articles' topics as closely related as possible. In order to be able to search and process the data, so that specific patterns can be identified, the data itself has to be transformed into a format that allows us to recognise those patterns. Such a transformation can be achieved by the application of a syntactic parser. The *Oc* makes use of the *Stanford parser* (Klein and Manning, 2003) to derive grammatical structures for each sentence, which then form a more accurate basis for the later pattern search. The *Stanford parser* is a freely available lexicalised PCFG (probabilistic context-free grammar) parser, that allows the user to employ a specific configuration. When extracted, the articles need to be transformed from their original HTML format to an appropriate sentence list representation. For this purpose, we use the *DocumentPreprocessor*[5]. After obtaining the articles and letting each sentence be processed by the parser, the *Oc* starts searching for specific patterns.

[4] http://j-spider.sourceforge.net/

[5] http://www.koders.com/java/

4 Experiments

More ambiguous patterns tend to introduce accuracy issues to the resulting ontologies and will compromise the ontolgies' reliability and thus also its usefulness overall. For this reason, it is necessary to separate more reliable patterns from those, which will be of very little value given a majority of ambiguous results. Thus, in section 4.1, we describe the evaluation of different lexico-syntactic patterns and afterwards discuss the most common errors that were observed. In section 4.2, we show how the respective successfulness/accuracy of a pattern influences the value of the ontology.

4.1 Pattern Evaluation

In order to evaluate a pattern's usefulness for automatic ontology creation, we assess each pattern's success rate individually. All patterns are tested on a corpus containing 733 *Wikipedia* articles (a sample consisting of 161585 sentences), that were collected across different areas to also ensure a pattern's applicability across genres. In the ideal case, patterns are both successful and frequent. A given lexico-syntactic pattern is considered to have matched correctly, if its results (hypernym/hyponym(s)) can be rephrased into a structure as the one presented in example (2).

(2) $NP_1,\ NP_2...,\ (and \mid or)\ NP_n\ is\ a\ NP_0$
$for\ all\ NP_i,\ 1 \leq i \leq n, hyponym(NP_i, NP_0)$

Hence, the following sentence: *"The other major European powers, such as the UK, still had high fertility rates..."*, which leads to the relation: *hyponym("UK","MajorEuropeanPower")*, needs to be rephrasable into: `UK is a MajorEuropeanPower`.

Another important point is the "one-directionality" of the respective pattern, meaning the position of hyponym and hypernym in relation to the pattern is not arbitrary. A match to a pattern should always display the same order of hyponym/hypernym: $hyponym\ (pattern - specific\ part)\ hypernym$, since otherwise processing can create false results.

The patterns used for the *Oc* (shown in table 1) were suggested by Hearst and Mititelu (Hearst, 1992; Mititelu, 2008). Some of the patterns were discarded for lack of results or performance reasons (more ambiguous patterns, such as the classic *IS-A* were not used here as the results were altogether too erroneous). Pattern grouping under the same number indicates a similarity in the pattern, that allows for a group search.

No.	Pattern
1.	NP_0 *including* NP_{1+i}
2.	NP_0 *such as* NP_{1+i}
3.	*by such* NP_0 *as* NP_{1+i}
4.	NP_0 *like* NP_{1+i}
5.	NP_0 *except* NP_{1+i}
6a.	NP_0 *e.g.* NP_{1+i}
6b.	NP_0 *i.e.* NP_{1+i}
7a.	NP_0*, (a) kind(s) | type(s) | form(s) of* NP_{1+i}
7b.	NP_0*: (a) kind(s) | type(s) | form(s) of* NP_{1+i}
8.	NP_0 *other than* NP_{1+i}
9.	*There (are | is) (could | would) be two types of* NP_0 *(: | ,)* NP_{1+i}
10a.	NP_0 *especially* NP_{1+i}
10b.	NP_0 *notably* NP_{1+i}
10c.	NP_0 *particularly* NP_{1+i}
10d.	NP_0 *usually* NP_{1+i}
10e.	NP_0 *mostly* NP_{1+i}
10f.	NP_0 *mainly* NP_{1+i}
10g.	NP_0 *principally* NP_{1+i}

Table 1: *Patterns for the acquisition of definitions*

No.	Overall occurrence	% of success	one-directional
1.	601	409 (68%)	No
2.	2389	2107 (88.2%)	Yes
3.	9	9 (100%)	Yes
4.	401	330 (82%)	Yes
5.	18	10 (56%)	Yes
6a.	170	134 (79%)	Yes
6b.	no occur.	nil	nil
7a.	48	31 (65 %)	Yes
7b.	7	6 (85%)	Yes
8.	19	16 (84 %)	Yes
9.	4	4 (100%)	Yes
10a.	61	9 (89%)	Yes
10b.	22	13 (59%)	Yes
10c.	29	23 (79%)	Yes
10d.	9	7 (78%)	Yes
10e.	5	4 (80%)	Yes
10f.	3	2 (67%)	Yes
10g.	no occur.	nil	nil

Table 2: *Pattern success rates*

Table 2 shows the results for the pattern evaluation. The number label indicates the specific pattern according to table 1. Column 1 displays overall occurrence in the whole corpus. Further, column 2 shows all successful ones out of all occurrences in both number and percent. The final column lists the directionality for each pattern. Only two patterns (1 and 2) obtained over 600 occurrences in the corpus. All others have results much lower than that. The most successful patterns (4 and 9), with a 100% accuracy, are also among the most infrequent. Patterns 6b and 10g did not occur at all in the used data.

4.1.1 Common Pattern Issues

In this section, we discuss the most common issues regarding the pattern search in our experiments.

Range A rather frequent issue is the one of range. Here, problems occur, when the extracted entities, NP_0 and NP_{1+i}, are not in a hypernym-hyponym(s) relationship, due to the fact that the hyponym(s) refer to another entity than the one extracted, as the pattern can vary in its scope. For example, let us consider the following sentence: *"Other foreign artists also settled and worked in or near Paris, like Vincent van Gogh..."* from which were extracted the relation: *hyponym("Vincent van Gogh","Paris")*, instead of the correct one: *hyponym("Vincent van Gogh", "ForeignArtist").*

One-directionality of a pattern Some patterns are not only one-directional. Thus, a match as the following is also returned: *"The newspaper created a new children section covering children books, including both fiction and non-fiction, and initially counting only hardback sales."*. Although, here "non-fiction/fiction children books" is implied, this match instead results in the relations: *hyponym("Fiction","ChildrenBooks") hyponym("NonFiction","ChildrenBooks").* The relation should be realised in the reverse order, as not all fiction or non-fiction books are children's books. If a pattern displays such a tendency, the latter can be particularly problematic, since even correct matches will produce incorrect results.

Pattern-specific issues An interesting case is the sentence: *"Lentil is also commonly used in Ethiopia as a stew like dish called Kik..."*. It shows a use of "like" other than in a construction, such as NP_0 *like* NP_{1+i}. Although the match does theoretically fit the pattern, its meaning does not entail the intended relationship of hyponymy.

Extra-embedded subclauses Another problem can be observed, namely that hypernym and hyponym(s) are not directly named after each other, but interrupted by a subclause, sometimes even containing another match as in *"There are two types of unsweetened cocoa powder: natural cocoa, like the sort produced by Hershey's and Nestlé using the Broma process, and Dutch-process cocoa, such as the Droste brand..."*

4.2 Ontology Creation

In the following part, we describe three different ontologies, that were created from the pattern matches. One is populated by the results for the patterns with an accuracy level of over 80%. The second features all remaining matches from the patterns with an accuracy of below 80%. The third combines all patterns. Since we cannot check the source for every relation in the ontology, we apply a more restrictive approach to the results. The aim is to determine the usefulness of the ontology overall. Therefore, it is only important, whether a relation in the ontology is correct and appropriate in terms of general content and the correctness of superclass/subclass relation.

Table 3 shows the results for the three ontology evaluations. In row 1 are the numbers for the more accurate patterns, below the less accurate ones and row 3 shows the combined ontology. The total number of the relations of the first setting is 4566, of which 3534 were correct and 1032 were incorrect. Respectively, the number of the relations from the second setting is 1508, of which 798 were found to be correct and 710 incorrect. The total number of the combined ontology with all patterns is 5823, of which 4140 were correct and 1683 incorrect.

Evaluation As this evaluation shows, there is little to be gained by using patterns with an accuracy of below 80%. Only 53.9% of the resulting ontology was correct. Whereas using more reliable patterns had an ontology accuracy of 77.4 %. For the ontology that used all patterns, an overall accuracy of 71.1% was achieved. Although, there is only about 6% difference between using only >80% accuracy patterns and using all patterns, this difference is mainly due to the fact, that the percentage of >80% patterns was overall much higher in the sample. Hence, using patterns with higher accuracy is likely to be effective in the long run. For simple class concepts, there are generally no problems. However, more complex concepts, as introduced by extra complements, do present difficulties in regard to scope, where a correct match will frequently be processed incorrectly. As *Wikipedia* is a relatively large resource, one may not have to rely on such problematic relations, since individual facts do occur more often. Another more general issue are "relational" words, such as: *different, related, nearby, comparable...* In most cases, these relate to a broader context and

setting	overall success	matched	successful match	unsuccessful match
1.	>80%	4566	3534 (77.4%)	1032 (22.6%)
2.	<80%	1508	798 (53.0%)	710 (47.0%)
3.	56-100%	5823	4140 (71.0%)	1683 (29.0%)

Table 3: *Ontology comparison*

bear less semantic relevance. It would therefore be worthwhile investigating their contribution to the ontology-building process. The question of making a new entity a class or an individual is also a complex issue, since there may be different semantic implications linked to it. Considering individual countries, there may be two possibilities; one is an interpretation for a country as an individual and the other as a class, which may have members itself: $\text{France} \in \text{Country} \bigvee (\text{Lorraine, Languedoc-Roussillon} \in \text{France}) \subset \text{Country}$.

For these reasons, also appropriate processing and representation of the results has to be considered.

Most errors are connected to issues as outlined in 4.1.1. Furthermore, it can be beneficial to analyse successful and frequent patterns more closely to see what grammatical constructions are most likely to occur in connection with them, so one can adapt the processing accordingly. As this rather superficial analysis is already able to effect a considerable increase in performance, looking at individual patterns in detail is reasonably worthwhile. In addition to using frequent and successful patterns, one can add the successful, but less frequent ones, as they do not put much strain on the system. Yet, their real accuracy level is questionable, since they do not occur often enough to confirm it.

5 Conclusion and Future Work

The overall aim of this project is to evaluate lexico-syntactic patterns in regard to their accuracy and reliability to match definitions in articles from online web sources, such as *Wikipedia*. Results are then transferred into an ontology representation using the language *OWL*. We show that using reliable patterns, one can create an ontology with an overall accuracy of 77%. However, some issues in connection to the incorrectly matched relations as well as processing remain. It is necessary to conduct larger experiments to find out how frequent a specific issue appears in text. Using lexico-syntactic patterns to extract definition relations has shown substantial success, which justifies a closer analysis of pattern ambiguity and other pattern-related issues. In general, since *Wikipedia* is a big resource with a vast amount of articles, one is able to afford losing some prospective facts for the benefit of precision and consequently to obtain a more accurate and thus also more useful knowledge base.

References

Ronald J. Brachman. 1983. What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, 16(10):30–36, Oct.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. ACL.

Marti A. Hearst. 1998. Automated discovery of wordnet relations. In *C. Fellbaum, WordNet: An Electronic Lexical Database*, pages 131–153. MIT Press.

Carmen Klaussner and Desislava Zhekova. 2011. Pattern-Based Ontology Construction From Selected Wikipedia Pages. In *Proceedings of RANLP 2011 Student Research Workshop*.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Lee W. Lacy. 2005. *Owl: Representing Information Using the Web Ontology Language*. Trafford Publishing.

Diana Maynard, Adam Funk, and Wim Peters. 2009. Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proceedings of WOP2009 collocated with ISWC2009*, volume 516. CEUR-WS.org, November.

Verginica Barbu Mititelu. 2006. Automatic extraction of patterns displaying hyponym-hypernym co-occurence from corpora. In *Proceedings of the First CESCL*.

Verginica Barbu Mititelu. 2008. Hyponymy patterns. In *Proceedings of the 11th international conference on Text, Speech and Dialogue*, TSD '08, pages 37–44, Berlin, Heidelberg. Springer-Verlag.