Acquiring Ontological Relationships from Wikipedia Using RMRS

Aurelie Herbelot and Ann Copestake

University of Cambridge, Computer Laboratory,
J. J. Thomson Avenue, Cambridge, United Kingdom
ah433@cam.ac.uk,ann.copestake@cl.cam.ac.uk

Abstract. We investigate the extraction of ontologies from biological text using a semantic representation derived from a robust parser. The use of a semantic representation avoids the problems that traditional pattern-based approaches have with complex syntactic constructions and long-distance dependencies. The discovery of taxonomic relationships is explored in a corpus consisting of 12,200 animal-related articles from the online encyclopaedia Wikipedia. The semantic representation used is Robust Minimal Recursion Semantics (RMRS). Initial experiments show good results in systematising extraction across a variety of hyponymic constructions.

Key words: ontologies, ontology extraction, Wikipedia, semantics

1 Introduction

Ontology extraction has traditionally relied on pattern matching algorithms. Hearst [5] introduced hyponymic extraction using lexico-syntactic patterns. In the Hearst algorithm, the system looks for instances of certain expressions in the text, for example 'X is a Y' or 'X such as Y and Z', and infers the relations 'X is-a Y' and 'X is-a Z'. Such systems are usually based on regular expressions over text or POS-tagged text, and sentences containing apposition, bracketing, long-distance dependencies or uncommon structures have to be coded explicitly. An example of such a sentence, extracted from the Wikipedia encyclopaedia, is:

> The Firemouth Cichlid is one of the 'typical' and most commonly seen (in pet stores) of the Cichlasoma-type South American cichlids.

Here obtaining the relationship 'Firemouth cichlid is-a cichlid' involves the identification of the hyponym and hypernym as the first and last noun phrases of a lengthy sentence. This suggests that traditional methods might be improved by using deeper syntactic and semantic analysis.

The work presented here investigates the use of a semantic model to address some of these issues. Robust Minimal Recursion Semantics (RMRS, [4]) provides argument-based representation of sentences. The theoretical idea is that, in the problematic sentence above, an RMRS output would contain the predicate associated with the identity copula *be* with a first argument corresponding to the

term *Firemouth cichlid* and a second argument corresponding to *cichlids*, regardless of the word order, modification and so on. Thus, having obtained the RMRS representation of a given corpus, it would be possible to extract ontological relationships from a semantic structure that abstracts over those morphological and syntactic details that do not affect the ontological relationship.

As pointed out by Pennacchiotti and Pantel [6], most ontology extraction systems so far have focused on generalised is-a or part-of relationships. Our work involves extracting general hyponymic relations with RMRS and applying a filter to the results to obtain biological, taxonomic relationships. The corpus was gathered by extracting 12,200 animal articles from the Wikipedia online encyclopaedia (`http://www.wikipedia.org/`), providing a semi-edited setting where the added robustness of semantics might prove its usefulness.

The next section of this paper gives an overview of relevant prior work. It is followed by the description of an extraction system based on RMRS, using hardwired rules. Results are discussed in the light of four different evaluation methods covering both manual and automatic recall and precision. A brief overview is then given of a further system, still under development, the aim of which is to automate the pattern extraction process. The conclusion presents different avenues for future work.

2 Previous Work

The most widely used framework for automatic ontology extraction was originally proposed by Marti Hearst [5], who introduced the use of lexico-syntactic patterns for the extraction of hyponymic relationships. Hearst's method has since then been followed by the most successful systems, such as the Espresso system proposed by Pennacchiotti and Pantel [6], based on bootstrapping.

Aside from this pattern-based approach, new clustering methods have also been investigated. The main idea is to group terms that appear in the same kind of context and label the resulting clusters. The process yields natural hyponymic relations between the members of the cluster and its label. The clustering approach was pioneered by Caraballo [2] who used conjunction and apposition to form new clusters.

Other methods attempt to do away with prior text processing on the basis that, as the size of the corpus increases, syntactic processing is not sustainable. Ravichandran and Hovy[7] propose a pattern matching algorithm for question answering systems, which relies on the direct use of the surface form of the corpus.

Our work takes an approach similar to those investigated by Hearst and Pantel: patterns containing a binary relation of interest are applied to the corpus to extract instances. However, the method uses prior syntactic and semantic processing at a deeper level than the techniques described above.[1]

[1] We are grateful to a reviewer for pointing out Suchanek et al's paper [8] (which appeared just after our work was completed). They show excellent results using a

3 Using RMRS for Ontology Extraction

3.1 An Introduction to RMRS

RMRS[4] is a development of Minimal Recursion Semantics [3]. One of its main features is its compatibility with both shallow and deep parsers, making it versatile enough for a wide range of applications: for the work reported here, we use RMRS structures derived from parses produced by RASP3 (Robust Accurate Statistical Parser [1]). RMRS allows for semantic underspecification and is robust in that a structure is produced even for partial parses. In the worst case, a (highly underspecified) RMRS can be constructed from POS-tagged data alone. Thus structurally deficient analyses can still yield correct ontological relationships as long as a connection can be found between the two elements of a pair.

For this work, we use a compiled form of RMRS in which each sentence in the corpus corresponds to a series of minimal trees. Each tree has a root, which is one of the lemmas in the sentence, and one or more daughters, the first one of which is the index of the lemma (the other daughters being potential arguments). As well as the argument trees, 'in-group' relations, denoted here by the tag 'INGS', express conjunction. The elements of the trees and in-group relations can be co-indexed to reconstruct the whole sentence or, if the complete parse is not available, phrases in the sentence. An example is shown in Fig. 1.

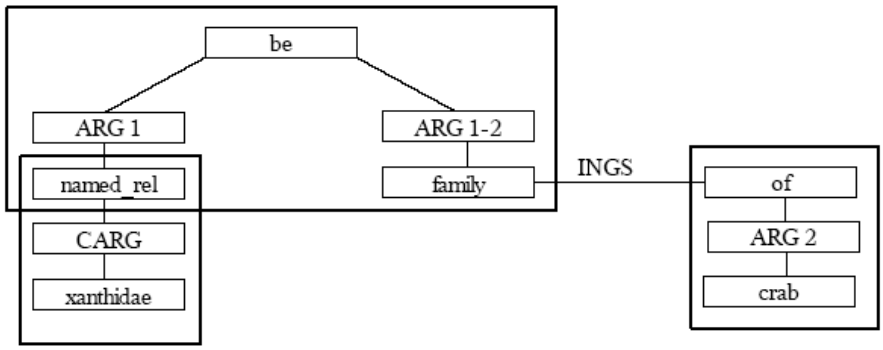

Fig. 1. RMRS representation for the sentence *Xanthidae is a family of crabs*. Indices are not shown. Three trees are identified by the rectangular frames.

parser of roughly similar depth to the one we use, but they rely on syntactically-labelled relations. In principle, the RMRS approach has the advantage that it allows more abstraction from irrelevant details of syntax and also that it is possible to switch between parsers without modification to the extraction code, and indeed to use merged results from more than one parser. See [4] for further discussion.

3.2 Corpus Preparation

The Wikipedia corpus is available as various database dumps in XML format. The dump used in this work was released on 22nd April 2006 and includes the most up-to-date versions of all pages, without edit history or pictures, at the date of the dump creation. 12,200 animal articles were extracted from the dump and preprocessed so that only text information remained. The resulting pages were parsed using RASP3 [1] and the RASP-RMRS converter [4] was applied to the derivations to obtain the RMRS.

The core of our system is a collection of classes that extracts and records RMRS trees, as described in §3.1, out of the original RMRS parse. The resulting representation forms the basis of ontology extraction.

3.3 System Design

An initial, manual annotation of 100 articles in the corpus showed that 50% of pages started with the simplest hyponymic relation: A is-a B, where A is the first argument and B the second argument of the identity copula. In order to test the capabilities of RMRS extraction, our system was designed to systematically process such relationships and a number of common variations such as *A is a species of B* or *A is a B species in the C family*. The patterns covered by the system are:

1. A is a B
2. A is a B species (or other taxonomic vocabulary)
3. A is a species of B (or other taxonomic vocabulary)
4. any of the above but with a 'one of' construct inserted prior to the hypernym (e.g., *the gecko is one of several species of lizards.*)

An assumption was made as to the directionality of the hyponymic relationship: the cases where the hypernym is indicated by the subject of the identity copula (*one of the most common species of felines is the cat*) were considered infrequent enough to bypass their separate treatment at this stage.

The patterns above required the identification of taxonomic vocabulary. Contributors to Wikipedia consistently follow the Linnaean scientific classification. There are seven levels in the Linnaean hierarchy (kingdom, phylum, class, order, family, genus and species); those levels were recorded in a vocabulary file for future use, together with a few variations and additional terms commonly seen in the articles' introductions.

The algorithm processes each corpus sentence in four steps and fully utilises the argument structure and relational blocks of the RMRS. For conciseness, the following gives the skeleton of the code only but provides at each step the list of RMRS features used.

```
1. Identify RMRS trees headed by the identity copula.
   ARG1 = base for hyponym
   ARG2 = base for hypernym
```

```
   Store ARG1 and ARG2 terms in arrays h and H.
2. Resolve 'one-of' constructs:
   with: one INGS of
   then: of ARG1 x (replace 'one' with x in H.)
3. Resolve taxonomic terms T:
   with: 'of' trees (of ARG1 T ARG2 resolution)
   or:  adjective trees (resolution ARG1 T)
4. Expand hyponym and hypernym:
   with: compound_rel, named_rel and adjective trees (JJ ARG1 NN).
```

The results of this process are potential hyponymic relations, which may refer to any kind of entity, not only animal organisms. A simple Named Entity Recognition method implemented by lexicon lookup was applied to the results to filer them. The NER lexicon was automatically constructed by recording the titles of all articles in the corpus. Truncation was systematically applied to both terms of the relation in order to extract terms such as *wild cat* out of the phrase *a small wild cat*.

4 Results and Evaluation

4.1 Evaluation Measures

Four evaluation measures were trialled, covering traditional calculations such as manual precision over a subset of the corpus as well as less common methods such as manual recall. The following describes all four heuristics. The Gold Standard used throughout the evaluation was the NCBI online taxonomy (`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy`).

1. **Manual recall.** Recall was calculated over a subset of the corpus. 100 articles were read by one annotator and each relevant hyponymy recorded, together with its originating sentence. 194 relationships were found and covered the use of syntactic patterns such as constructs depending on the identity copula, apposition, punctuation and full inclusion verbs (*include*, *comprise*, etc). The manual recall score is thus the number of unique pairs extracted from the subset divided by 194.[2]
2. **'Rough' recall.** This recall figure reflects how many pairs were found in relation to the number of articles considered. Using the figures compiled for manual recall calculation, it was estimated that each article contained roughly two unique taxonomic relationships. The rough recall was simply the number of unique pairs divided by twice the number of articles.
3. **Manual precision evaluation.** Manual precision was calculated over 100 pairs randomly extracted from the results. The aim was to check the actual truth of the extracted relations rather than just their correspondence to the originating sentences. The scoring scheme was therefore designed to take into account the reliability of the sources used in checking:

[2] Due to time constraints, evaluation was performed by one annotator only: we hope to remedy this in future work.

   (a) If the pair is found in the NCBI database or on the site of a scientific/academic organisation, score +1
   (b) If the pair is found on any other site, score +0.5
   (c) If the pair is not found, score 0.

   Wikipedia mirrors and Wikipedia itself did not count as sources.
4. **Automatic precision evaluation.** Manual annotation is very time-consuming, but the repeated use of a very small set of manual annotations for system development carries a high risk of over-fitting. So a program was written to compare all pairs in the results with the NCBI database. The number of pairs found gives a 'minimal precision' figure which was used in devlopment to investigate the relative precision of different versions of the system.

4.2 Results

Running the initial system on the whole corpus resulted in 3985 relations being extracted, giving a 16.5% 'rough' recall figure. Manual recall yielded a 14% figure. RMRS gave promising results in the treatment of complex structures and long distance dependencies. Some examples of the relationships extracted by our system are shown below:

1. The Cottontop Tamarin (Saguinus oedipus), also known as the Pinchu Tamarin, is a small New World monkey weighing less than 1lb (0.5 kg): *cottontop tamarin is-a new world monkey*
2. The Norway lobster, Nephrops norvegicus (also called Dublin Bay prawn or langoustine), is a slim orange-pink lobster found in the north-eastern Atlantic Ocean and the Mediterranean Sea: *norway lobster is-a lobster*
3. Opah (also known colloquially as moonfish, sunfish, kingfish, and Jerusalem haddock) are large, colourful, deep-bodied pelagic Lampriform fish comprising the small family Lampridae (also spelt Lamprididae): *opah is-a fish*

The precision, as calculated over 100 randomly chosen pairs was 92% – the errors and partially verified pairs (scored as 0.5) are shown below:

Pairs	Comments	score
bear is-a spectacled bear	Incorrect order	0
lamnidae is-a great white shark	Incorrect order	0
nemegtbaatar is-a djadochtatherioidea	Non-academic website	0.5
rayonneceras is-a cephalopod	Non-academic website	0.5
bat is-a falcon	Incorrect truncation	0
tiger is-a leopard	? Perhaps incorrect parse	0
kootenia is-a trilobite	Non-academic website	0.5
pin-tailed whydah is-a songbird	Songbird not a taxonomic type	0
jaguarundi is-a wild cat	Wrong analysis of *wild cat*	0
engrailed is-a moth	Non-academic website	0.5

Out of six incorrect pairs, two resulted from the inverted 'hypernym is-a hyponym' pattern mentioned in §3.3 not being catered for. As non-experts, we

cannot be completely certain whether the Wikipedia entries are correct or not, since even if we find supporting evidence on another site, we cannot be sure that site is correct. But these results show no evidence of unreliability in Wikipedia. The use of *wild cat* in the jaguarundi article to refer to species other than domestic cats and big cats is perhaps an infelicity, given that it also refers to a specific species, but is not an error.

Applying the automatic evaluation to the results yielded a score of 44%, i.e. less than half of the relations could be verified against the NCBI database. Inspection of the evaluation output file revealed that 1757 hypernyms were not found in the NCBI names list. This was indicative of the potential usefulness of ontology extraction in this domain, even though large-scale ontologies already exist.

4.3 Evaluation

The main problem with our system is obviously the low recall. Several factors contribute to explain this result:

1. Some correct pairs are lost when checking the terms against the animal lexicon (the list is not comprehensive enough).
2. Incomplete parses affect the final result.
3. The number of extraction rules is small. Expanding the rules to include verbs such as *encompass*, *comprise*, etc should provide a broader basis for extraction.

Lexicon Issues In order to ascertain the shortcomings of the lexicon, the first 250 pairs found by the system were printed out prior to any term matching being performed, i.e. non-taxonomic pairs were part of the set. A manual check revealed that out of those 250 pairs, 34 taxonomic pairs were already being extracted by the system but 23 additional pairs contained possible valid taxonomic names as their hyponym and hypernym, yielding a potential 67% increase on the number of relations extracted in that particular section of the corpus.

There were four main reasons for relations being lost during list lookup:

1. The search term did appear in Wikipedia but under a variant (for instance, *theropod* is not an article but *theropoda* is. Hence only *theropoda* was listed in the lexicon.)
2. The term did not appear in Wikipedia at all.
3. The parser had not transformed a plural noun into its singular equivalent, leading to a search failure.
4. Compounds were being extracted with the wrong word order.

An attempt was made at remedying to the lexicon's shortcomings using Wikipedia redirections. It was hoped that recording all animal redirections would provide a comprehensive list of the noun variants used in the encyclopaedia. Over 1M redirection pages were extracted from the dump and 21,531 animal entries

recorded. The system was expanded with a redirection checking stage and plural noun processing. The compound issue was solved by ensuring that terms were concatenated in their order of appearance in the RMRS. The modified program yielded 4771 pairs, or a 20% increase on the original figure, pushing the rough recall figure to 20%. Manual precision dropped slightly to 88.5%.

Partial Parses Issue An inspection of the RMRS file for the subset used at manual recall stage showed that many instances of 'be' in that part of the corpus were contained in partial parses and missing argument values: out of 194 potential pairs, 97 depended on the identity copula. Out of those 97, 41 were missing an argument altogether or an argument value, indicating that only about 29% of the manually extracted pairs were recoverable. This issue would be partly solvable by expanding the set of rules responsible for converting the syntactic parse into an RMRS parse. The other avenue to explore is the syntactic parse itself. This work only considers the first parse obtained from RASP. Sometime errors occur in parts of the parse that indicate the taxonomic relationship. This might be addressed by processing the n-best parses from the RASP output, rather than just the first, although there are cases where the correct parse is not found at all, or has very low rank. In the longer term, we intend to investigate the use of deeper parsing in conjunction with RASP.

Rule Coverage This initial system is restricted by the limited number of patterns considered during extraction. The next section describes how we have started implementing a new system that automatically extracts RMRS rules out of a training set and applies them to our corpus to discover new instances. This new design ensures that patterns are kept separate from the extraction module and shows potential in increasing recall by using a wider variety of rules.

5 A First Attempt at Automatic Pattern Extraction

5.1 Defining the RMRS Pattern

Let A and B be two English terms. A and B are known at training stage (they are the training instances) and unknown at pattern matching stage (they are the new pairs extracted by the system). The RMRS pattern is the path linking A and B through the RMRS representation of the sentence. A path is an ordered set of unique coindexed RMRS trees. Each RMRS tree consists of either a lemma with all its identifiable arguments or an INGS relation. No loop is allowed in the path. An example is shown below:

```
LEMMA::be ARG1::A ARG2::member|LEMMA::member ARG1::group|
LEMMA::compound_rel ARG1::group ARG2::B
```

The pattern in the example consists of three trees. The ARG2 of the first is co-indexed with the lemma of the second. The ARG1 of the second is co-indexed with the ARG1 of the third. A is the hyponym and B is the hypernym.

The hyponymic relationship may correspond to one or more patterns, depending on whether it contains compounds or not. Relational patterns (the core of the hyponymy) and compound patterns are extracted and stored separately. So for instance, the training pair 'yellow-bellied elaenia is-a bird' leads to a relational pattern linking 'elaenia' and 'bird' and a compound pattern linking 'yellow-bellied ' and 'elaenia'. Two assumptions were made here:

1. The last word of a compound, and the last word only, constitutes the hyponym or hypernym's core.
2. In the case of multiple element compounds (*Northern Brown Bandicoot*), each adjective/qualifying noun modifies the main noun, i.e. the last word in the compound. (This is a simplification.)

5.2 Initial Experiments

Using the definition given above, the first iteration of our new system was run over a training set consisting of the 194 pairs manually obtained from the corpus when performing manual recall calculations. 32 relational patterns and 24 compound patterns were extracted. These figures are small compared to the size of the training set, but error evaluation shows that the extraction failures are due to issues that could be fixed in further work. The incomplete RMRS rule set is once again the main problem. In particular rules were not available to treat apposition of the type *The Northern Brown Bandicoot, a marsupial species, is a bandicoot...* or noun plus qualifier constructions of the type *the family Drepanidae*. Wikipedia articles frequently omit commas round appositions, which prevented their recognition. Slight tuning or retraining of the parser should improve performance.

The experiments also demonstrated the necessity for a targeted or sufficiently large training set. The rules we extracted in these preliminary experiments did not cover all potential taxonomic terms such as *family*, *genus* or *species*. Some over-specific lexical items were also observed on the patterns' paths.

In order to get an indication of the potential in RMRS pattern extraction, we replaced extracted taxonomic terms with the generalised dummy entry 'taxovoc' in the rules and changed all over-specific lexical entries into general equivalents. We then run a pattern matching algorithm over the resulting rules. This yielded a large increase in recall with 9142 unique pairs being extracted: a 37% rough recall. Manual precision, however, was only 64.5% while the automatic minimum precision figure fell to 30%. A systematic rule evaluation showed the presence of highly imprecise patterns, highlighting the need for a thorough automatic rule evaluation method in further iterations of the system development.

6 Conclusion and Further Work

This work attempted to move away from the traditional regular expression-based approaches in ontology extraction by using a semantic framework. It was

shown that RMRS offers advantages in terms of tackling structural complexities. Applying a systematic treatment of the identity copula's arguments to a corpus consisting of 12,200 Wikipedia articles resulted in 4771 taxonomic relationships being extracted: roughly 20% of the available pairs, yielding 88.5% precision.

These promising initial results call for the development of a more complete system, including pattern extraction and pattern matching features. Some trials with automatic rule extraction were reported here, showing a large potential for recall increase balanced with a need for a thorough pattern evaluation method.

Further work will focus on implementing such a system and will attempt to resolve the issues highlighted in this paper. In particular, it is expected that expanding the set of syntactic-to-semantic conversion rules in the RMRS and running the system on different syntactic parses might produce a significant improvement in recall without affecting the precision. Other domains (chemistry being a possible choice) will also be investigated in order to show whether RMRS can be used in a general purpose tool.

Acknowledgments This work was supported by the UK Engineering and Physical Science Research Council (EPSRC: project EP/C010035/1). We are also grateful to the anonymous reviewers for their comments.

References

1. Briscoe, E. and Carroll, J.: Robust Accurate Statistical Annotation of General Text. In: Proceedings of the Third International Conference On Language Resources and Evaluation (LREC 2002). Las Palmas, Canary Islands (2002)
2. Caraballo, S.: Automatic Acquisition of a Hypernym-Labelled Noun Hierarchy from Text. In: Proceedings of ACL-99. Baltimore, MD (1999) 120–126
3. Copestake, A, Flickinger, D, Sag, I and Pollard, C.: Minimal Recursion Semantics: an introduction. In: Journal of Research on Language and Computation 3(2–3) pages 281–332 (2005)
4. Copestake, A.: Robust Minimal Recursion Semantics. `http:www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf`
5. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: COLING-92. Nantes, France (1992) 539–545
6. Pennacchiotti, M. and Pantel, P.: A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations. In: Proceedings of Inference in Computational Semantics (IcoS-06). Buxton, England (2006) 87–96
7. Ravichandran, D. and Hovy, E.: Learning Surface Text Patterns for a Question Answering System. In: Proceedings of ACL-2002. (2002) 41–47
8. Suchanek, F., Ifrim, G. and Weikum, G.: LEILA: Learning to Extract Information by Linguistic Analysis. In: Proc 2nd Workshop on Ontology Learning and Population (at ACL-2006), pp 18–25, Sydney, 2006.